

Steering the future of computing

Computational power is surging thanks to insatiable consumers. Natural scientists should seize opportunities to stimulate computer science, to help everybody cope with huge volumes of data.

Sometime in the 2010s, if all goes well, the Large Synoptic Survey Telescope (LSST) will start to bring a vision of the heavens to Earth. Suspended between its vast mirrors will be a three-billion-pixel sensor array, which on a clear winter night will produce 30 terabytes of data. In less than a week this remarkable telescope will map the whole night sky with a greater speed and sensitivity than could have been imagined more than a decade or so ago, recapitulating with added detail the entire history of optical astronomy from Galileo to the Palomar Sky Survey.

And then the next week it will do the same again, looking for transient changes, adding new information and building up a database of billions of objects and millions of billions of bytes.

When looking at the future of scientific computing, as *Nature* does this week in a selection of News Features and Commentaries (starting on pages 398 and 409), it is easy to focus on the vast data architectures necessary for projects such as the LSST or the Large Hadron Collider at CERN, the European particle-physics laboratory near Geneva. The truly amazing story, though, is of the distributed power that ends up not in exceptional places such as the focal plane of a giant telescope, but spread out across the world; the power that allows data to be acquired from microfluidic chemistry sets and genome sequencers in labs around the world at astonishing rates, and allows the environment — or the human body — to be monitored in real time by vast arrays of sensors. The fact that everyday computing is getting exponentially cheaper promises to vastly increase data flows of all sorts and to revolutionize the practice of science.

It is this remarkable growth that has allowed projects such as the LSST to be imagined — and which will surpass them before they are very old. It is not driven by science, but it has been of immense use to scientists, and will continue to be, if they can change the way science is done to make use of the great potential.

Scientists will increasingly have to rely on automation to extract useful knowledge from these vast data resources. As with computer-aided proofs in mathematics, such automation challenges the

processes by which scientists gain insight and generate theories. What's more, science will increasingly be done directly in the database, finding relationships among existing data while someone (or something) else performs the primary collecting role. And this means that scientists will have to understand computer science in much the same way as they previously had to understand mathematics, as a basic tool with which to do their jobs.

But scientists can be more than just passive responders to change. Although the great trends in computing are driven by economic and technical forces external to the scientific world, science can provide ideas and challenges that provoke the computer industry into moves it might not have made so quickly on its own. The World Wide Web, after all, grew out of the needs of scientific data users. It was years after Tim Berners-Lee had put his vision of hypertext onto the Internet that it revealed its capacity to revolutionize fields from bookselling to campaign financing.

"Science can provide ideas and challenges that provoke the computer industry into moves it might not have made so quickly on its own."

The computer industry knows that scientists can come up with strange ideas and requirements that may well, in time, have broader commercial application elsewhere. This is one of the reasons why Microsoft is engaging the scientific community with its new 'Towards 2020 Science' report on computers in science (see <http://research.microsoft.com/towards2020science>). That report inspired this week's focus on computing in *Nature*. Microsoft is sponsoring free web access to our articles on the subject, although, as always, the content is exclusively *Nature's* responsibility.

As computing gets ever cheaper, quicker and more powerful, scientists would do well to remember that, by being a demanding and stimulating 'user community' that engages the interest of companies such as Microsoft, Google and Intel, they can influence the development of the field, to everybody's benefit. ■

A scramble for Africa

Large dams benefit contractors and corrupt governments more than they aid the African people.

Towards the end of nineteenth century, Europe suddenly woke up to the riches that lay in the vast unexplored continent to its south, and the 'scramble for Africa' began. By the start of the First World War, almost all of the continent had been taken by European powers. The rights of Africa's own people, who lost land and many lives during this process, drew scant attention.

Why recall this episode today? Fleeting, last summer, Africa was big news, when it became the central topic at a meeting in Scotland of the leaders of the G8 group of top industrialized nations, chaired by British prime minister Tony Blair. Yet the real action is being taken by a donor nation that isn't even a member of the G8: China.

The G8 nations — correctly, if belatedly — are considering how best to invest in Africa, so that the previous misappropriation and mismanagement can be avoided. China seems to have no such qualms. Across the continent, from Zimbabwe to Sudan, China is winning friends by lending money to Africa's most unsavoury regimes without asking awkward questions.

As a News story on page 393 of this issue illustrates, scientists and



engineers are sometimes complicit with this process. Sudan's Merowe dam on the Nile could be set to repeat the mistakes that have characterized previous large-scale hydropower projects in poor countries. Studies of how to resettle 50,000 people whose land will be flooded, and assessments of the project's environmental impact, were finished late in the day, and undertaken with insufficient rigour. They would have stopped the project going ahead if the World Bank, for example, was funding it.

Lahmeyer International, the German company that is coordinating the project, is disarmingly open about why this is so. It says that funders such as the World Bank make things too complicated. Thorough environmental and social impact assessments take years; Sudan wants power now. China is willing to invest, in part to cement closer ties with an important oil producer. And the Sudanese government lacks the political infrastructure — and probably the political will — to enforce proper safeguards. So, once again, thousands of poor people look set to suffer so that a big dam project can go ahead.

The project's backers have sought to portray Merowe as a necessary trade-off between the competing needs of development and the rights of local people. But there is no reason why both needs can't be met. Hydropower certainly has a role to play in Africa's development. Most of the continent's available hydropower resources are untapped and could, if properly harnessed, provide a valuable and renewable source of energy. But that doesn't mean that large dams need to be built. Successful projects in Asia and South America have shown that small hydropower projects can supply a few thousand

local people without the need for big resettlement projects. Smaller projects can be run with more input from local people and are easier to combine with other renewable sources, such as solar power.

Unless the lessons of the past are thoroughly learned, large dam projects will sink over time in a morass of corruption, haphazard displacement of local people, lack of political accountability, and failure to plan properly for maintenance.

South Africa has, to its credit, tried to incorporate some of these lessons into a hydropower and water-supply project in Lesotho. The project is imperfect, but at least its administrators have sought to consult with local people and to run independent assessments of its environmental impact. But South Africa, with its wealth and its relatively sophisticated political system, is an exception.

In many other African nations, there is little chance of proper safeguards being implemented. Chinese firms and government agencies will operate with few checks or balances. The same goes for the European companies involved in Merowe and elsewhere. They know that rigorous political consultation and environmental assessment are needed if big dam projects are to succeed. Yet they have been happy to engage in such projects in the absence of any such safeguards. The staff and shareholders of these firms are part of another scramble for Africa, in which local peoples' rights and needs are once again being sidelined in the stampede for wealth. ■

"People's rights and needs are once again being sidelined in the stampede for wealth."

A colourful past

The production of dyes in the nineteenth century marked a turning point in the appliance of science.

The 150th anniversary of William Perkin's synthesis of aniline mauve dye (see page 429) is more than just an excuse to retell a favourite story from chemistry's past. To be sure, the tale contains much to delight in: Perkin's extraordinary youth and good fortune, the audacity of his gamble in setting up a business to mass-produce the dye, and the chromatic riches that so quickly flowed from an unpromising black residue of coal-gas production. But perhaps the most important aspect of the story is the relationship that it engendered between pure and applied science.

The demand for new, brighter and more colourfast synthetic dyes, along with new means of setting them on to fabrics ('mordanting'), stimulated manufacturing companies to set up their own research divisions, and thus cemented interactions between industry and academia that were just developing at the time.

Traditionally, dye-making was a craft, a combination of trial-and-error experimentation and the rote repetition of time-honoured recipes. The idea that chemical production required real scientific expertise did not arise until the eighteenth century, when the complexities of mordanting and multicolour fabric printing moved beyond the expertise of mere recipe-followers.

That was when the Scottish chemist William Cullen announced that if the mason wanted cement, the dyer a dye and the bleacher

a bleach, "it is the chemical philosopher who must supply these". Making inorganic pigments preoccupied some of the greatest chemists of the early nineteenth century, notably Nicolas-Louis Vauquelin, Louis-Jacques Thénard and Humphry Davy. Perkin's mauve, however, was an organic compound and so, in the mid-nineteenth century, was rather more mysterious than metal salts. The drive to understand the molecular structure of carbon compounds during this period is often presented today as 'pure' chemistry, but in reality it owed much at the time to the profits that might ensue if the molecular secrets of organic colour could be unlocked.

Both the need to understand molecular structure and the demand for synthetic methods were sharpened by chemists' attempts to synthesize indigo and alizarin (the natural colour obtained from the madder plant). When Carl Graebe and Carl Liebermann found a route to making alizarin in 1868, the Badische dye company, soon to become BASF, quickly acquired the rights. One of those who found a better route in 1869 was Ferdinand Riese, who was already working for Hoechst. Another was Perkin.

These and other dye companies, including Bayer, Ciba and Geigy, had seen the value of having highly skilled chemists on their payroll — something that was even more evident when they branched into pharmaceuticals early in the twentieth century. Thus, today's integration of scientific research into industry first began to take shape, as companies realized that good business needs good scientists. ■

"Dye companies, including Bayer, Ciba and Geigy, had seen the value of having highly skilled chemists on their payroll."

RESEARCH HIGHLIGHTS

Blowing its top

Geophys. Res. Lett. **33**, L05313 (2006)
A newly identified type of ash-laden plume from volcanic eruptions may cause distant destruction, raising troubling issues for disaster planning.

A study of the 2002 Reventador eruption, 100 kilometres from Quito, Ecuador, found an unusual composition of steam and cool ash in the plume. The plume could then unexpectedly deposit damaging pyroclastic ash flows far from the volcano.

Pinaki Chakraborty and his colleagues from the University of Illinois in Urbana-Champaign suggest that the scallop-shaped plume (pictured) of the Reventador eruption was caused by fluid instability from the negative buoyancy of plume materials.



A. ALVAREZ SANCHEZ, CRUZ ROJA ECUATORIANA/AGU

EVOLUTION

Under the spotlight

Science **311**, 1617–1621 (2006)

The molecular components of the lizard's third, or parietal, eye — a light-sensitive spot thought to sense changing light conditions — reveals something about how light detection evolved in the two familiar vertebrate eyes.

Working with side-blotched lizards (*Uta stansburiana*; pictured below), Chih-Ying Su and King-Wai Yau of the Johns Hopkins University in Baltimore, Maryland, and their colleagues identified one opsin protein most sensitive to blue light, and another most sensitive to green light, in the same cells of the third eye. The researchers show that the two proteins use different intermediary G proteins to respectively close and open ion channels in the cell.

This light-detection machinery may retain

more features from ancestral vertebrates than is seen in the rods and cones of modern vertebrate eyes, the authors propose.

CANCER

Deadly decisions

Cancer Cell **9**, 157–173 (2006)

High-grade gliomas (HGGs), a particularly aggressive type of brain tumour, are essentially incurable and kill most patients within months. Now Heidi Phillips at Genentech in South San Francisco, California, and her colleagues have identified genome-wide expression profiles and signalling pathways for a large cohort of HGGs, which may lead to more effective therapies for the disease.

Activating different signalling pathways that are crucial for controlling normal brain development gave rise to varied subtypes of HGGs — each with its own prognosis and disease progression, the researchers found.

Interestingly, the subtypes of tumours have gene-expression profiles of cells at various stages of development. This suggests that tumour growth might be regulated by mechanisms that regulate cell-fate decisions during neurogenesis.

BIOLOGY

Timing is everything

Curr. Biol. **16**, 512–515 (2006)

Wild hummingbirds can remember precisely where and when to find their food, a new study suggests.

Laboratory experiments indicate that some animals can learn short time intervals

between feedings. The new work, from Susan Healy of the University of Edinburgh, UK, and her colleagues, supports the notion that wild animals can do the same — specifically in their quest to find nectar.

In tests, rufous hummingbirds (*Selasphorus rufus*) learned to distinguish between a set of four flowers that were refilled with sugar solution 10 minutes after being emptied, and another set of four that were refilled 20 minutes after being emptied. The birds' timing ability is more detailed than previously thought, the researchers write.

IMMUNOLOGY

Not-so-innocent bystander

J. Exp. Med. doi:10.1084/jem.20052056 (2006)

Once thought to have only a role in autoimmunity, natural regulatory T cells do double duty by influencing immune responses to both self and foreign antigens, researchers have found.

Yasmine Belkaid, of the National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, and her colleagues infected mice with the parasite *Leishmania major*. They found that natural regulatory T cells replicated rapidly at the primary infection site and attacked the microbe but not the mouse. These cells also contribute to the bystander effect, in which other immune cells in the area are suppressed, thus complicating infection control.

Regulation of this effect could lead to improved vaccines for diseases such as cancer, AIDS and malaria, where these cells have been found to play a part.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

ASTROPHYSICS

Turn them on, dead neutrino

Phys. Rev. Lett. **96**, 091301 (2006)

Sterile neutrinos — neutrinos that do not easily interact with others — could account for most of the Universe's 'dark matter', according to a hypothesis by Peter Biermann of the Max-Planck Institute for Radioastronomy in Bonn, Germany, and Alexander Kusenko of the University of California, Los Angeles.

The authors show that sterile neutrinos, each with a mass of a few kilo-electron-volts, could explain some other odd features of the Universe. For example, sterile neutrinos ejected from supernovae could kickstart the fast-spinning stars known as pulsars. And in the early Universe, X-rays from their decay might accelerate star formation.

CELL BIOLOGY

High pressure

Cell **124**, 929–942 (2006)

Narrowing of the arteries causes high blood pressure, a very common condition whose molecular basis is poorly understood.

Researchers in Italy now demonstrate an intriguing interaction between the signalling molecule TGF- β — which is critical for blood-vessel development and blood-pressure control — and Emilin1, a protein secreted by the extracellular matrix.

Mice whose *Emilin1* genes were deleted had high blood pressure. Emilin1 reduces levels of TGF- β by interacting with the molecule proTGF- β , from which TGF- β is normally generated. This mechanism of TGF- β regulation is novel and may have more general implications, as TGF- β is involved in many other pathological processes including cancer.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

NEUROBIOLOGY

Astrocytes' star role

Neuron **49**, 823–832 (2006)

Astrocytes, the star-shaped glial cells in the brain (pictured above), help stimulate formation of the insulating myelin sheath around nerve fibres, according to a study led by Douglas Fields at the National Institute of Child Health and Human Development in Bethesda, Maryland.

Electrical stimulation triggers nerve fibres to release ATP, which serves as an important signal for myelin to form. Surprisingly, the researchers showed that the ATP did not directly act on the myelin-forming cells in the central nervous system, known as oligodendrocytes. It acted on astrocytes, causing them to secrete a regulatory protein known as LIF, which in turn promoted the myelinating activity of oligodendrocytes.

BIOCHEMISTRY

Take the heat

J. Am. Chem. Soc. **128**, 3144–3145 (2006)

Most proteins don't like heat, a fact that can hinder their use in biotechnology and industrial chemistry. Warm them up and

their chains tend to unfold. How might they be made more robust?

Daniel Raleigh and his colleagues at the State University of New York at Stony Brook suggest that, instead of trying to stabilize the folded protein, one might aim to destabilize the unfolded, or denatured, state. They introduced into the protein fragment NTL9 two mutations that were predicted to discriminate against the unfolded state: one cuts out a favourable electrostatic interaction, and the other lowers the entropy.

Whereas pristine NTL9 is already partly unfolded in a urea solution at room temperature, the double-mutant version survives at least up to 60 °C.

GENETICS

A gut sense of ageing

Cell **124**, 1055–1068 (2006)

Regulators of lifespan that convey signals from the gonad to the gut have been identified in genetic studies in the nematode worm *Caenorhabditis elegans*.

Cynthia Kenyon of the University of California, San Francisco, and her colleagues previously found that removing the cells that produce sperm and eggs extends lifespan. In these worms, a FOXO-family transcription factor moves into the nucleus of gut cells.

But how do the cells in the gonad relay signals to the gut? Kenyon's team now identifies three proteins that could be responsible. One (KRI-1) may help to usher the transcription factor into the nucleus. Another (DAF-12) seems to operate upstream as a nuclear hormone receptor that responds to a hormone produced by the third protein (DAF-9).

Proteins similar to all three of these factors have been identified in mammals.

JOURNAL CLUB

Anthony R. Fiorillo
Dallas Museum of Natural
History, Texas

**This palaeontologist uses
lizards from down south to make
sense of dinosaurs up north.**

We, as humans, can regulate our body temperature through internal means. This ability, known as endothermy, is a marvellous thing. However, endotherms typically need an insulating layer of fur or

feathers to stay warm. (This is why parents engage in endless battles with their children over the need to wear a coat in the cold.)

Some other creatures, such as reptiles, are ectotherms — meaning they use ambient thermal sources to regulate body temperature.

In a recent article, Richard Shine of the University of Sydney in Australia explored how ectothermy may have shaped the evolution of reptiles' life history (*R. Shine Annu. Rev. Ecol. Evol. Syst.* **36**, 23–46; 2005).

He argues that the ability of reptiles to decouple the time of energy acquisition (feeding) and energy expenditure (reproduction) helps them adapt to different environments. It allows reptiles, for example, to withstand months of starvation. This intrigued me because it might bear on my own research.

High above the Arctic Circle in northern Alaska, I have been working with a team that has excavated thousands of dinosaur bones. And our work suggests that the animals lived in the Arctic all

year round. The climate then was much warmer than it is today, but the dinosaurs still had to contend with long, dark winter seasons. We don't think they had fur and it is unclear which dinosaurs had feathers, so how did they survive?

The decoupling of energy flow could be the key. Arctic dinosaurs may have starved during the winter and reproduced during the short summer. The trend is to think that dinosaurs were endotherms, but perhaps they were reptiles — extreme examples of the ectothermic approach to life.

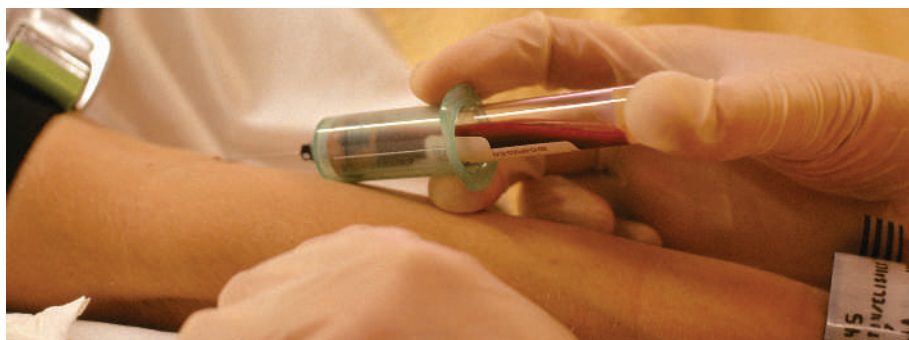
NEWS

London's disastrous drug trial has serious side effects for research

A drug trial that took a shocking turn in London last week may have far-reaching effects on policy. Its failure, experts say, could change restrictions on clinical research and increase scrutiny of the private companies that carry out the majority of clinical trials.

"There's going to be a lot of soul searching," says Thomas Murray, a bioethicist at the Hastings Center, a think tank in Garrison, New York.

As *Nature* went to press, two previously healthy young men were in critical condition and another four seriously ill at Northwick Park Hospital in London. On 13 March, they received intravenous injections of TGN1412, an antibody made by Boehringer Ingelheim for TeGenero, a small, privately owned biotechnology firm in Würzburg, Germany. The drug was being developed to fight autoimmune diseases and leukaemia. Parexel International, a contract research organization based in Waltham, Massachusetts, that operates in 39 countries, was running the trial for TeGenero.



IMAGESTAT/ALAMY

"Was informed consent adequate? Were the right subjects selected? Were the right doses given? This better have been done right, or some tough questions are going to come up for the private, commercialized research sector," says Arthur Caplan, a bioethicist at the University of Pennsylvania, Philadelphia.

Within one to two hours of being injected, the six volunteers suffered violent reactions that included headache, backache, nausea, a drop in

blood pressure and, ultimately, multiple organ failure. The trial was the first test of the drug in humans; it was immediately suspended by the UK Medicines and Healthcare Products Regulatory Agency, which is now investigating.

Thomas Hanke, chief scientific officer for TeGenero, says that the company had "no pre-clinical evidence whatsoever" that the drug might be unsafe, and that no adverse effects had been observed in rabbit and monkey studies.

THE DRUG TEST: WHAT WENT WRONG?

What is the drug?

TGN1412 is a monoclonal antibody being developed by the German company TeGenero. It is designed to direct the immune system to fight leukaemia, or relieve joints inflamed by rheumatoid arthritis.

How does it work?

The antibody stimulates a subset of immune cells called T cells by binding to a receptor molecule on their surface called CD28 — a pivotal molecule in the immune system. Normally, a T cell needs two incoming signals before it is activated: one from CD28 and a second from a separate T-cell receptor. TGN1412 is unusual because it overrides this basic control mechanism: when it binds to CD28, it activates the T cell without the need for a second signal.

What could explain the patients' reaction?

Doctors do not yet know, but multiple organ failure could be due to a contaminant in the medicine, the wrong dosage or the drug acting on the immune system in an unexpected way. Immunologists suspect that TGN1412 may have triggered T cells to release a toxic flood of molecules called cytokines. Or perhaps T cells attacked the body's own tissues, ignoring the safety mechanisms that normally keep them in check.

Are similar drugs already in use?

There is one, called anti-CTLA4 antibody, which binds to a different receptor on T cells and boosts the effects of the CD28 receptor, albeit to a lesser degree. Scientists working with this drug have seen side effects in human trials — including skin rashes

and gut reactions, some of which can be controlled with steroids. Anti-CTLA4 antibody is now in late-stage clinical trials to treat certain cancers.

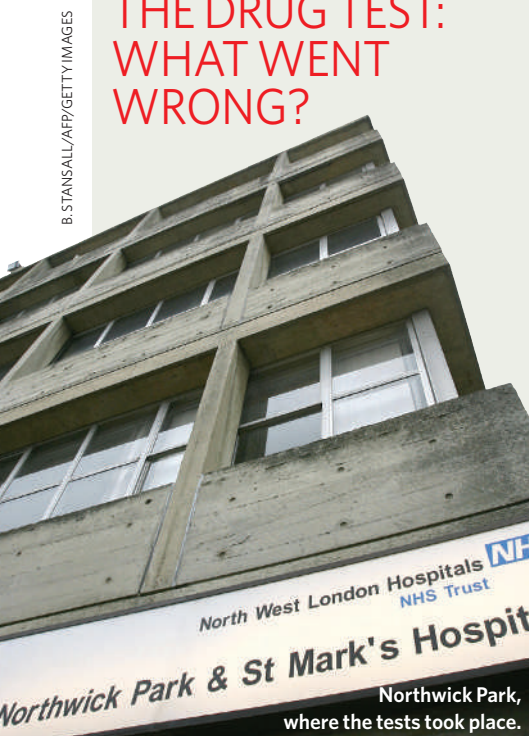
Were there problems during animal tests with TGN1412?

TeGenero says that it saw no problematic side effects in tests on rabbits and monkeys. But researchers know that an antibody designed to bind to and rouse human CD28 may not work in the same way as the slightly different animal versions of the same molecule.

Will the drug be abandoned?

Immunologists are divided. Some warn that the approach is inherently dangerous. Others hope that TGN1412 or drugs like it might still find a use if researchers can learn to control its side effects or target specific tissues.

Helen Pearson



B. STANSALL/APP/GETTY IMAGES



SHIP ENDURES RECORD-BREAKING WAVES
Storm shows extreme seas may be more common than thought.
www.nature.com/news

But some observers say the company should have been more cautious, because the drug aimed to bypass the immune system's natural control mechanisms (see "The drug test: what went wrong?"). "You are going beyond the regulatory network, so all hell can break loose," says Angus Dalgleish, an immunologist at St George's Medical School in London.

Reaction to the trial's outcome has been swift, and mixed. In an unexpected turn of events, Britain has seen more people enquiring into early trials, in which healthy volunteers take experimental drugs to determine their safety and dose range. "Paradoxically, there's been an upsurge in interest in these healthy volunteer studies," says Max Parmar, who heads cancer studies at the Clinical Trials Unit of the UK Medical Research Council. The men who were struck ill in the TeGenero trial had been paid £2,000 (US\$3,500) each to do the test.

In Germany, reaction has been rather different. The local public prosecutor in Würzburg is investigating whether any criminal wrongdoing was involved. And this week, the Paul Ehrlich Institute, which authorizes human trials of biological drugs, announced it will tighten regulation of the first tests of such drugs in people. "The central question is: why do you treat six people at the same time? Why don't you start with one?" says Johannes Löwer, president of the Langen-based institute.

Löwer says that his institute will start requiring sequential rather than simultaneous administration of 'high risk' monoclonal antibodies — those that, like TGN1412, activate central pathways of the immune system. He says that one-at-a-time administration — with days between injections — will also be required for experimental biologics if institute scientists are not convinced by animal-model studies, or no appropriate animal model exists.

Ethicists in the United States, meanwhile, have called for careful scrutiny of a newly loosened set of rules for the making and testing of drugs in early human trials (see page 406). And the \$10-billion business of contract research organizations, or CROs, has come under the microscope.

Caplan worries that such organizations are tacitly encouraged not to focus on protection for human subjects. "The CROs are often told: 'Just get us the data on the deadline.' They don't get asked questions about how that's being done." The Association of Clinical Research Organizations boasts that CROs conduct clinical trials 30% more quickly than the pharmaceutical companies that hire them.

A spokesman for Parexel says: "We believe that best practices were followed and the appropriate policies and procedures were adhered to."

Meredith Wadman

GRAPHIC DETAIL

GAS CENTRIFUGE

Magnetic Suspension Bearing
Top Scoop
Molecular Pump
Outer Casing
Rotor
Bottom Scoop
Stator
Bottom Bearing
Inverter

Indian Rare Earths Ltd.
408, Marathi Commercial Centre,
P/18-A, D. Anil Desai Road, West, Mumbai-400 028, INDIA.
Fax: 91-22-2496 0508 Phone: (022) 5890 4782-5
TENDER Advertisement No. IREADV196
Indian Rare Earths Ltd., Mumbai, India

Aluminium tubing has many uses, but these dimensions suggest that it could be for the outer casing of a centrifuge.

TP-719 Rs. 400/- 28.2.2006
Seamless Aluminium extruded tubes of 195 mm O.D. and 25 mm wall thickness as per specification

TP-719 Rs. 500/- 17.12.2003
Precision machining of special steel Thin Rod-R(L) with spherical tip.

TP-672/TPT Rs. 2000/- 25.1.2006 1030 hrs. 1130 hrs.
Manufacture and supply of Ø 150 mm x 1225 mm long Thin Wall Flow Formed Interstage Annealed Ultra Precision Tubes made out of MDN 350 grade free issue extruded preforms as per specification.

This advert may be for a supercritical centrifuge rotor assembly — one of the parts that is most difficult to build. An outside agent posing as a bidder might have been able to obtain far more detailed specifications from Indian Rare Earth.

Are adverts revealing nuclear secrets?

WASHINGTON DC

When US President George W. Bush was asked recently whether he thought India — with whom he had just announced a deal to export nuclear technology — was a responsible nuclear nation, he responded simply: "I do." But critics say a scan through the local papers is all it takes to show that New Delhi is blatantly circumventing US and European export controls, and publicizing nuclear secrets.

In a report released on 10 March, David Albright and Susan Basu of the Institute for Science and International Security in Washington DC reveal that a Mumbai-based company has frequently placed advertisements in Indian newspapers; the ads are for corporations to supply items that seem to be components of a uranium gas centrifuge.

They analysed almost 200 public advertisements placed since 1984 in *The Times of India* by Indian Rare Earths, a mineral-extraction company that they suspect is helping to further the government's uranium-enrichment programme.

Gas centrifuges are spinning canisters that can be used to produce uranium that is enriched in the fissile isotope uranium-238, for use in reactors or weapons. The technology is currently being pursued by nations such as Iran (see *Nature* 432, 432-437; 2004).

Not everyone is in agreement over how to respond to the ads. Arjun Makhijani, president of the non-profit Institute for Energy and Environmental Research based in Takoma Park, Maryland, points out that India never signed the treaty on the non-proliferation of nuclear weapons. So it is technically allowed to further its enrichment programme. "India is not breaking any rules," he says.

But Albright says the technically detailed descriptions being published in newspapers show India's lax control over its nuclear technology. "It's gotten ridiculous," he says. "India just can't turn its back on this, and neither can the United States."

Geoff Brumfiel

ON THE RECORD

“I’m sorry if I’m making people a little frightened, but I feel it’s my role.”

Virologist Robert Webster warns that bird flu could kill half the world’s population.

“On shuttle missions we often see mosquitoes... They seem very confused and die very quickly.”

A space-station astronaut comments on creatures that are unwittingly launched into space.

Sources: ABC News, ARRL

SCORECARD

**Flying saucers**

Patent-spotters unearth a 1973 British Rail patent that seems to be for a futuristic craft running on ‘thermonuclear fusion’. Sadly, the project never made it off the ground.

**Compost canard**

Entomologists squelch the rumour that gardening mulch made from trees felled by Hurricane Katrina is packed with destructive Formosan termites from the area.

**Sandy snow**

Dust storms in Northern China cause yellowish snow to fall across South Korea. Health officials warn that breathing in the sandy flakes might have ill effects.

OVERHYPED

The Tunguska blast

The meteorite that exploded above Siberia in 1908 has been credited with felling 60 million trees over 2,000 km². Now it is being blamed for global warming. Vladimir Shaidurov of the Russian Academy of Sciences says that the impact may have disturbed the distribution of atmospheric water vapour and increased global temperatures. But the big blast wouldn’t have done this, many climate scientists counter. Among other objections, they argue that the localized explosion could not have kicked off an unstoppable change in the dynamics of Earth’s atmosphere.

Trauma trials leave ethicists uneasy

SAN DIEGO

A US trial of an experimental blood substitute given to trauma patients who cannot give consent is stirring concern about the way that such ‘no consent’ trials are run.

It is rare for experimental treatments to be given without consent. In the United States, for example, the practice was authorized in limited circumstances only a decade ago. But such trials are set to increase under an initiative to test treatments for trauma and heart-attack victims. Advocates say that the tests are the only way to gather critical information in emergency situations — but critics argue that patients’ rights are not being sufficiently protected.

In the blood-substitute study, communities are supposed to be educated about the programme in lieu of each individual giving prior consent. But ethicists worry that, in at least some of the areas involved, consultation has been inadequate and there is little public knowledge about the risks, and how to avoid taking part.

Initiated in 2003, the study seeks to enrol 720 patients as they are rushed to 32 hospitals nationwide, with about 600 people having been enrolled so far. The oxygen-carrying blood substitute, called PolyHeme, is made by Northfield Laboratories in Evanston, Illinois. It consists of treated haemoglobin from human blood, and is designed to replace the saline solution given by emergency workers to prevent shock from blood loss.

Some ethicists are particularly concerned about blood transfusions being withheld from

patients once they are hospitalized, to see how well the blood substitute works. “I have a lot of problems” with various aspects of the work, says attorney Nancy King of the University of North Carolina in Chapel Hill.

King is one of three medical ethicists who last week wrote an open letter criticizing the trial’s construction to the *American Journal of Bioethics*¹. They are concerned that the trial conflates conditions in ambulances, where blood transfusions are not a feasible part of patient care, with conditions in hospitals, where they are.

Northfield officials refused interview requests, directing enquiries to the company’s website, which states that “Northfield is committed to conducting its study with the utmost concern for patient safety”.

Over the past year, ethicists have become increasingly vocal about various issues relating to the trial. One problem is that the product is often provided in poor communities with high proportions of ethnic minorities, as these can have higher numbers of trauma cases.

The only way to opt out of receiving PolyHeme if injured in areas covered by the trial is to wear a particular blue wristband. But critics say there have been inadequate advertisements and education meetings, and interviews with members of the emergency services and

IMAGE
UNAVAILABLE
FOR
COPYRIGHT
REASONS

T. STEWART/CORBIS

Chemists shrug off unseemly spotlight

The decision by Dalibor Sames to withdraw two published papers^{1,2} and part of a third³ has drawn unwanted attention to the chemical field of C–H functionalization. But the furor seems unlikely to dim the area’s lustre.

Sames says in his retraction⁴ that his group at Columbia University in New York has been unable to reproduce the published results since graduate student and co-author Bengü Sezen left the lab. Sezen stands by the results and says that she is prepared to repeat the work under Sames’s supervision. Columbia University, meanwhile, has launched an investigation into the matter.

C–H functionalization — the art of replacing carbon-bound hydrogen atoms in organic molecules with something more interesting — is unlikely to be badly damaged. “There are dozens, even hundreds, of exciting papers published every year,” says Alan Goldman of Rutgers University in Piscataway, New Jersey, author of a recent survey of the field. “I don’t think the retractions will cast any shadow.”

Bonds between carbon and hydrogen atoms are ubiquitous in the raw materials from which synthetic chemists make new molecules. Unfortunately, it is a serious headache to cut the strong bond between



Are the rights of trauma patients being eroded?

community liaisons suggest that the level of awareness of the trial is very low.

Another concern is a lack of information about the trial's results. The study is funded by Northfield, which pays the hospital \$10,000 per enrolled patient, and the company doesn't want the details disclosed to protect its financial interests. In San Diego, one of the study sites, Northfield unsuccessfully sued the *Reader* newspaper in December to block disclosure of details of the study. The newspaper is itself suing the University of California, San Diego (UCSD), which has enrolled about 25 trauma patients, for information about what it believes are adverse study results. UCSD officials say they are protecting patient privacy and abiding by Northfield's contract.

Northfield is also under fire for not promptly disclosing the results of a previous study that tested PolyHeme in patients with abdominal aortic aneurysms. On 22 February *The Wall Street Journal* reported that 10 of 81 patients suffered heart attacks — with two dying — after receiving PolyHeme. The study was halted in 2001.

Tong Gan, a clinical researcher at Duke University in North Carolina, and other principal investigators on the study have argued for the full results to be published so that scientists can learn from them. Northfield says it did not delay publication and that PolyHeme wasn't responsible for the problems.

But the results prompted at least four hospitals, including Albany Medical College in New York, to temporarily suspend the trauma study. Some have since restarted the trial. "Any institution that continues this trial without additional consultation with communities is doing a profound disservice," says Glenn McGee, director of the Alden March Bioethics Institute at Albany.

Michael Caligiuri, UCSD's director of clinical research programmes, counters that those involved with the trial are doing their best to deal with the ethical issues as they arise. "This is not a perfect system," he says. "We are learning a lot from the PolyHeme study."

In the coming months, a National Institutes of Health (NIH) research project called the Resuscitation Outcomes Consortium (ROC) is likely to spur similar questions as it starts enrolling patients. This \$50-million, five-year project at ten hospitals in the United States and Canada will test experimental techniques on trauma and heart-attack patients who cannot give prior consent. The first such study, of a high-salt saline solution for trauma patients, is set to begin later this year.

Paediatric cardiologist Tracey Hoke, the NIH's medical officer for the ROC, insists that lessons from the PolyHeme trial will have been learned: "We intend to fully disclose risks and benefits to the communities." ■

Rex Dalton

1. Kipnis, K., King, N. M. P. & Nelson, R. M. *Am. J. Bioeth.* doi:10.1080/15265160600668644 (2006).

a particular hydrogen atom and the carbon to which it is attached and replace that hydrogen with something else: another carbon, say, or a phenyl group. Most techniques will indiscriminately break all such bonds in any given molecule.

The work at the Sames lab and elsewhere aims to replace this blunt approach with something more delicate: employing reusable catalysts to fashion reactions targeted at specific bonds. Such selectivity would have practical applications in industrial processes to make pharmaceuticals or fuels.

Chemist Robert Bergman at the University of California, Berkeley, is sanguine about the technique breaking out of the lab. "If you asked people ten years ago whether anyone would ever come up with a

catalytic method to do this, they would have said no. I don't think it is outrageous to say that in five or ten years there will be commercial applications."

"This should have hit the news because it was right and exciting, instead of hitting the news because it was wrong," laments Travis Williams, a postdoc in the field at Berkeley. "I am sorry that the world is going to think that chemists get it wrong, because, almost always, chemists get it right." ■

Emma Marris

1. Sezen, B. & Sames, D. *J. Am. Chem. Soc.* **126**, 13244–13246 (2004).
2. Sezen, B. & Sames, D. *J. Am. Chem. Soc.* **127**, 5284–5285 (2005).
3. Godula, K. Sezen, B. & Sames, D. *J. Am. Chem. Soc.* **127**, 3648–3649 (2005).
4. Sames, D. *J. Am. Chem. Soc.* **128**, 3102 (2006).

SPECIAL REPORT



S. TORFINN/PANOS

Up to 50,000 people will lose their homes to the Merowe dam.

Tide of censure for African dams

A wave of Chinese-built dams in Africa, particularly the Merowe project in Sudan, could have devastating consequences for local communities. **Jim Giles** reports.

On the arid plains that surround the Nile north of Khartoum in Sudan, a huge dam is slowly taking shape. The billion-dollar Merowe project will more than double the amount of electricity that Sudan can produce, and is just one of a dozen new dam projects being built across Africa using Chinese money and expertise.

But scientists and environmentalists who have studied the dam say that poor local people will suffer because necessary precautions are not being taken. According to the first independent review of the dam plans, a copy of which has been seen by *Nature*, inadequate thought has been given to the environmental and social consequences of flooding hundreds of square kilometres of land. That is far from unusual when it comes to Chinese investment in Africa, environmental groups allege. They say that China, which has a dire domestic environmental record, is repeating the mistakes of previous big dam projects, and that rural African communities will pay the price.

"Chinese companies will ignore social and environmental impacts to the extent that local governments are willing to ignore them," says

Thayer Scudder, an anthropologist at the California Institute of Technology in Pasadena, who has spent decades studying hydropower projects. "If governments don't care, or are corrupt, why will the Chinese engineers worry?"

The surge of large dam projects that began in the 1960s caused many lessons to be learned the hard way. Projects in China, India, North America and elsewhere caused serious downstream erosion through the removal of sediment, and the majority of resettled people suffered a decline in quality of life (see 'Lessons to learn', overleaf). Hydrologists say that thorough assessments of the impact, good design and well-funded resettlement programmes can minimize the impact of new dams.

Large scale

But although several national governments and the World Bank have earned praise for implementing some of these ideas, experts say China has not followed suit. At Merowe, for example, where commercial electricity production is scheduled to begin next year, a 7-kilometre-long, 65-metre-high wall will trap more than 10 million cubic metres of water. Land belonging

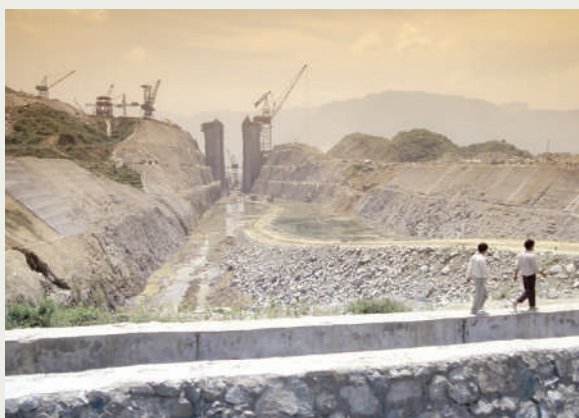
to 10,000 families will be affected, and four resettlement schemes are under way. A project of this scale in Europe or North America would have required years of independent study before being given the go-ahead, say dam experts. But the first environmental-impact assessment was completed in 2002, just a year before construction began.

Contrary to rules for dam projects adhered to by other big funders such as the World Bank, the assessment was conducted, not by independent experts, but by Lahmeyer International, the German company acting as engineering consultants for the project. The assessment, which was never formally published in Sudan, also fails to tackle key issues. It falls well short of international standards, say researchers at the Swiss Federal Institute of Aquatic Science and Technology in Kastanienbaum, who will publish their review of the assessment on 23 March.

Sediment accumulating behind the dam is one such issue, says Cristian Teodoru, a geologist at the institute who produced the review. This will affect farmers downstream who rely on nutrients in the sediment to fertilize their

Lessons to learn

- The Aswan High Dam in Egypt, built across the Nile during the 1960s, caused severe erosion downstream. Sediment that would normally have replenished the area remained trapped behind the dam, and salt water from the Mediterranean eventually inundated parts of the delta.
- Few recent hydropower projects have been as controversial as China's Three Gorges Dam (pictured). The US\$25-billion dam, which will create a lake 600 kilometres long and is due for completion at the end of the decade, drew particular criticism over the forcible resettlement of more



than half a million people.

- The author Arundhati Roy has led a high-profile campaign against dam building in India, in particular the Narmada Dam Project in central India. The

Indian government has several major dams under construction and in planning, but Roy and others say previous projects caused severe damage to local ecosystems and communities.

fields during the annual floods. "This issue is not even mentioned in the report," says Teodoru. The future of downstream ecosystems gets similarly short shrift. Teodoru says that the water column at the front of the dam will contain layers of anoxic water, which, if sent through the turbines and released, could harm aquatic life below the dam — another point the assessment fails to acknowledge.

Egon Failer, an executive director at Lahmeyer in Bad Vilbel, defends the report, saying that it was an early-stage document and that the final design addresses some of the problems that Teodoru identified. Dam gates will, for example, be opened to flush out sediment in July when the floods arrive and stir up water in the reservoir. He adds that the centre of the turbine inlet is just 12 metres below the dam surface, so that oxygen levels will be high enough not to harm downstream ecosystems.

Inaccessible

But details of the final design have never been published, and scientists working for Lahmeyer were required to sign confidentiality agreements that prevent them from talking to the media, making it impossible for researchers and environmental groups to assess Failer's assertions. Teodoru points out that even 12 metres below the surface oxygen levels could be dangerously low. Water levels will also fluctuate during the year, and could leave the base of the turbine twice as deep as Failer indicated. Many modern projects minimize the problem of anoxic water by using designs that extract water from different depths, or incorporate devices that mix the water to prevent anoxic areas forming, says Karin Seelos, a

senior environmental adviser at Hydro-Québec, an energy company in Montreal, Canada.

Critics of Merowe say the project also looks set to repeat perhaps the most severe social mistake of previous dam schemes, which have seen the livelihoods of hundreds of thousands of resettled people suffer. Mutaz Musa Abdalla Salim, director of financing at the Merowe Dam Project Implementation Unit in Khartoum, denies this. He points out that 35% of the 10,000 affected families have been resettled and US\$700 million will eventually be spent on new homes, clinics and other facilities. But when staff at the International Rivers Network, an environmental group in Berkeley, California, which part-funded the Swiss study, visited the resettlement sites in February 2005, they catalogued many problems. The group say sanitation was already poor owing to cramped conditions in the new towns, and that soil quality was in some places too low to support agriculture. Salim angrily denied these assertions when *Nature* put them to him.

Experts are concerned that the problems identified at Merowe will affect other Chinese-built dams. Press reports suggest that China is considering, or has already started work on, at least a dozen dam projects in Africa. These projects tie in with investments in many other African projects, some in countries with poor human-rights records, in a strategy that Western political commentators say is designed to secure access to the continent's oil and other natural resources.

Chinese government officials did not return requests for comment on their country's

involvement with Merowe or other African dams. But Jia JinSheng, vice-president of the China Institute of Water Resources and Hydropower Research in Beijing, confirms that his country is winning contracts in the region. He says Chinese firms can undercut rivals in Europe and North America by around a third, and finish construction more quickly. He adds that environmental and social issues have become more important in China in the past five years, but that ultimately responsibility rests with the host nation: "It's the business of the owners and local government," he says.

Local responsibility

Those parties are often ill-equipped to provide the necessary safeguards, says Ute Collier, an expert on dams at the WWF in Godalming, near London. Collier published a report on hydropower in Africa on 6 March in association with UK charities WaterAid and Oxfam. She backs the potential of hydropower, pointing out that around a third of the continent has no access to electricity. But she says that only a few African countries, such as Zambia and South Africa, have the political will and infrastructure to ensure that hydropower is implemented responsibly. Elsewhere, she says, funders with low environmental standards, such as China, or the Arab banks that are also contributing to the Merowe project, can operate almost unchecked.

Pressure could potentially come from the third group involved in dam building: the commercial firms that run the projects. But

"If African governments don't care, or are corrupt, why will the Chinese engineers worry?"

Collier says that some Western companies seem happy to work on such projects. As well as Lahmeyer, two other European firms are involved in Merowe: the Swiss company ABB, which sells equipment for electricity sub-stations, and Alstom in France, which is supplying turbines and generators. "Companies should operate in the same way as they do at home," says Collier. "But that is often not the case."

On the issue of resettlement, for example, Lahmeyer seems to have set aside its own reservations about the project. The company's 2002 assessment acknowledged that, with just over a year to go before construction, there was no agreed plan on how to resettle the 50,000 people involved and that prospects of "smoothly" resettling at least one group — members of the Shaigiya tribe — were "doubtful".

"Maybe the assessment should have been done earlier," says Failer. "But the country didn't have the time. They need electricity. The water is there and they're pushing for it. I'm deeply convinced this is good for the country." ■



DID EARTH SEED LIFE ELSEWHERE IN THE SOLAR SYSTEM?

Impacts on our planet could have sprayed life into space.

www.nature.com/news

Microwave data refine picture of Universe

You probably wouldn't expect measuring the Universe to be easy. But turning data from the Wilkinson Microwave Anisotropy Probe (WMAP) into a map of the Big Bang's afterglow has been an all-consuming task.

"We just had to get this out to get on with our lives," says Lyman Page of Princeton University, New Jersey. Page and his colleagues have spent the past three years analysing the latest data from WMAP, a satellite launched in 2001 to study patterns

in the radiation known as the cosmic microwave background. These patterns, which reflect the state of the Universe at the time the first atoms formed, hold clues to its structure, history and composition.

The team's results, revealed on 16 March, broadly confirm its first conclusions — presented in 2003 after analysis of a single year of data.

The 2003 data provided powerful support for the idea that the Universe was dominated by things not directly observable — dark matter and dark energy. To improve on these results, which were based on maps of temperature fluctuations, the team mapped the polarization of the microwaves. The temperature signals are tiny (between 30 and 70 millionths of a kelvin); the polarization

signal is 100 times fainter still.

To detect the polarization signal meant characterizing completely the noise in the instrument. "At times you just thought, oh my gosh, how are we ever going to get this right?" says Gary Hinshaw, who coordinated the analysis.

For example, an amplifier on the satellite introduces complex noise into the data by contaminating each observation with a lingering trace of its previous measurement. The team had to find a way to correct for this across the 3 million pixels of the full-sky map. Often, they would try an approach only to find that it introduced new artefacts.

"It was just gruesome detail day after day," says Charles Bennett, WMAP's principal investigator.

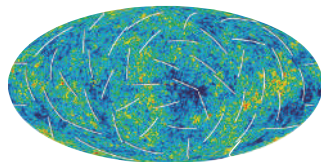
From the polarization maps, the

team precisely calculated the Universe's 'optical depth', one of the six fundamental properties of the cosmos. This allowed it to estimate that the first stars formed 400 million years after the Big Bang.

It calculated another basic property, known as the inflation parameter, to be 0.95 with an uncertainty of less than 0.02. Theories of inflation, which hold that the Universe had a dramatic growth spurt in its first moments, predict that the inflation parameter should not be equal to one.

The results, submitted to *The Astrophysical Journal*, were released four days before the WMAP team was due to submit a request for further funding to NASA. The work goes ever on.

Jenny Hogan



Hotspots: afterglow of the Big Bang.

NASA/WMAP SCIENCE TEAM

Board of Canadian journal sides with sacked editors

The editorial board of the *Canadian Medical Association Journal* jumped ship last week. Of its 19 members, 15 resigned, saying they do not trust the leadership of the Canadian Medical Association (CMA).

The journal's top two editors were sacked last month (see *Nature* 440, 10–11; 2006). Board members believe it was because they published items at odds with the CMA's views on contraception and private health care.

The association responded with an open letter accusing the old editorial board of “undermining” the journal. The CMA has recruited Antonio Lamer, retired chief justice of the Supreme Court, to review the relationship between the association, its subsidiary CMA Holdings and the journal.

“The journal is in crisis,” says ex-board member Philip Devereaux.

Cosmologist Barrow bags prize for divinity

Noted mathematician and cosmologist John Barrow of the University of Cambridge, UK, has won this year's Templeton Prize,



Thanks for everything: John Barrow (right) is honoured for his work on divinity and infinity.

worth £795,000 (US\$1.4 million).

Created by philanthropist John Templeton, the prize recognizes discoveries that “expand human perceptions of divinity”. Previous recipients of the prize include physicist Charles Townes, cosmologist George Ellis, Mother Teresa and evangelist Billy Graham.

Barrow's work includes more than 15 books, over 400 scientific papers and a popular play. He draws on mathematics, physics and astronomy to explain the behaviour of the Universe, theories of everything, and such mysteries as infinity and nothingness.

Science committee chair makes retirement plan

The powerful chairman of the US House of Representatives science committee, Sherwood Boehlert (Republican, New York), announced last week that he will not seek re-election next year. He had already planned to stop chairing the science committee, as required after three terms.

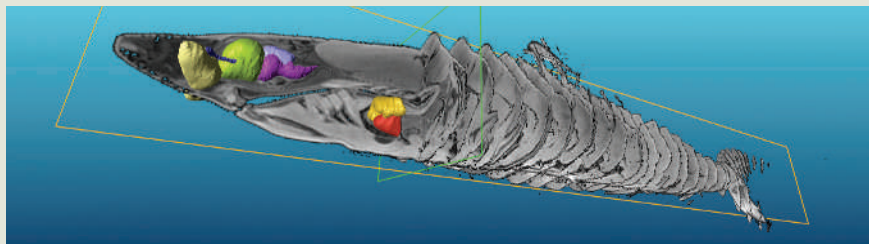
In his 24 years in Congress, including six as head of the science panel, Boehlert has played a key role in government debates over science, including US participation in the International Space Station and the international fusion experiment ITER.

Speculation is rife over who might succeed Boehlert. Committee member Bart Gordon (Democrat, Tennessee) is a likely candidate if the Democrats regain control of the House after next year's elections, and there are many potential Republican successors.

Retraction costs Hwang his licence and his job

Woo Suk Hwang of Seoul National University in South Korea lost his licence to conduct embryonic stem-cell research on

D. KARP/AP



Fish images provide net benefit to researchers

US researchers are creating a Digital Fish Library containing anatomical images of about 1,000 fish, such as the Galapagos shark pictured above, and will make it available on the Internet.

Scientists at the University of California, San Diego, will use functional magnetic resonance imaging to scan fish specimens representing about 500 species from the collection at the Scripps Institution of Oceanography.

A new 'digital dissection tool' for exploring the high-resolution images will be made available to researchers and science teachers. The catalogue of three-dimensional scans is to be completed in five years under a \$2.5-million grant from the National Science Foundation.

16 March, and was fired by the university on 20 March. There are now no groups licensed to carry out such research for therapeutic cloning in South Korea, although several are licensed for *in vitro* fertilization-based human embryonic stem-cell research.

The Ministry of Health and Welfare's decision to revoke Hwang's licence followed the retraction of his two papers on human

embryonic stem cells published in *Science*. But the retractions meant that Hwang no longer met the legal requirement of having at least one publication in embryonic stem-cell science in the past three years.

The decision came ten days after Hwang admitted ordering researchers to fake data for his 2005 *Science* paper, and nine weeks after the National Bioethics Committee released a damning report on the methods

used by Hwang to obtain eggs for his work.

Results of a criminal investigation into Hwang's research and use of funds are expected in early April.

Solid support for helium's second quantum state

Helium, already known for its bizarre 'superfluid' behaviour, has been confirmed to have a second, strange quantum state.

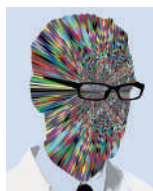
Three independent teams have now seen 'supersolidity' in helium-4, a phenomenon first reported by Moses Chan and Eun-Seong Kim of Pennsylvania State University two years ago (*Nature* 427, 225–227; 2004).

Supersolidity appears when solid helium is cooled below 0.3 kelvin and rotated slightly. Through quantum effects, some of the solid then begins to flow.

But results presented at the American Physical Society's meeting in Baltimore, Maryland, last week have spawned confusion about the physics underlying the phenomenon. For example, experiments led by John Reppy of Cornell University suggest that the effect disappears if the crystal is heated and then slowly cooled to remove defects. But Chan says this is at odds with his team's experience.

CHAMPING AT THE BITS

Despite some remaining hurdles, the mind-bending and frankly weird world of quantum computers is surprisingly close. **Philip Ball** finds out how these unusual machines will earn their keep.



Five years ago, if you'd have asked anyone working in quantum computing how long it would take to make a genuinely useful machine, they'd probably have said it was too far off even to guess. But not any longer.

"A useful computer by 2020 is realistic," says Andrew Steane of the quantum-computing group at the University of Oxford, UK. David Deutsch, the Oxford physicist who more or less came up with the idea of quantum computation, agrees. Given recent theoretical advances, he is optimistic that a practical quantum computer "may well be achieved within the next decade".

This excitement is, however, tempered by the hurdles that have yet to be overcome. Building a quantum computer is still very, very hard to

do. This is partly because it involves making quantum systems do things that don't come naturally to them. "There is progress, but it's still very slow," says physicist Chris Monroe of the University of Michigan in Ann Arbor.

And even if we did have a working quantum computer today, there are hardly any programs that could run on it. In fact, it is likely that even once the machines are available, quantum computers are destined to remain niche products — excellent for certain tasks but not versatile devices like conventional personal computers. "Quantum computers will almost certainly never become general-purpose desktop machines," concedes Isaac Chuang, a quantum physicist at the Massachusetts Institute of Technology (MIT) in Cambridge.

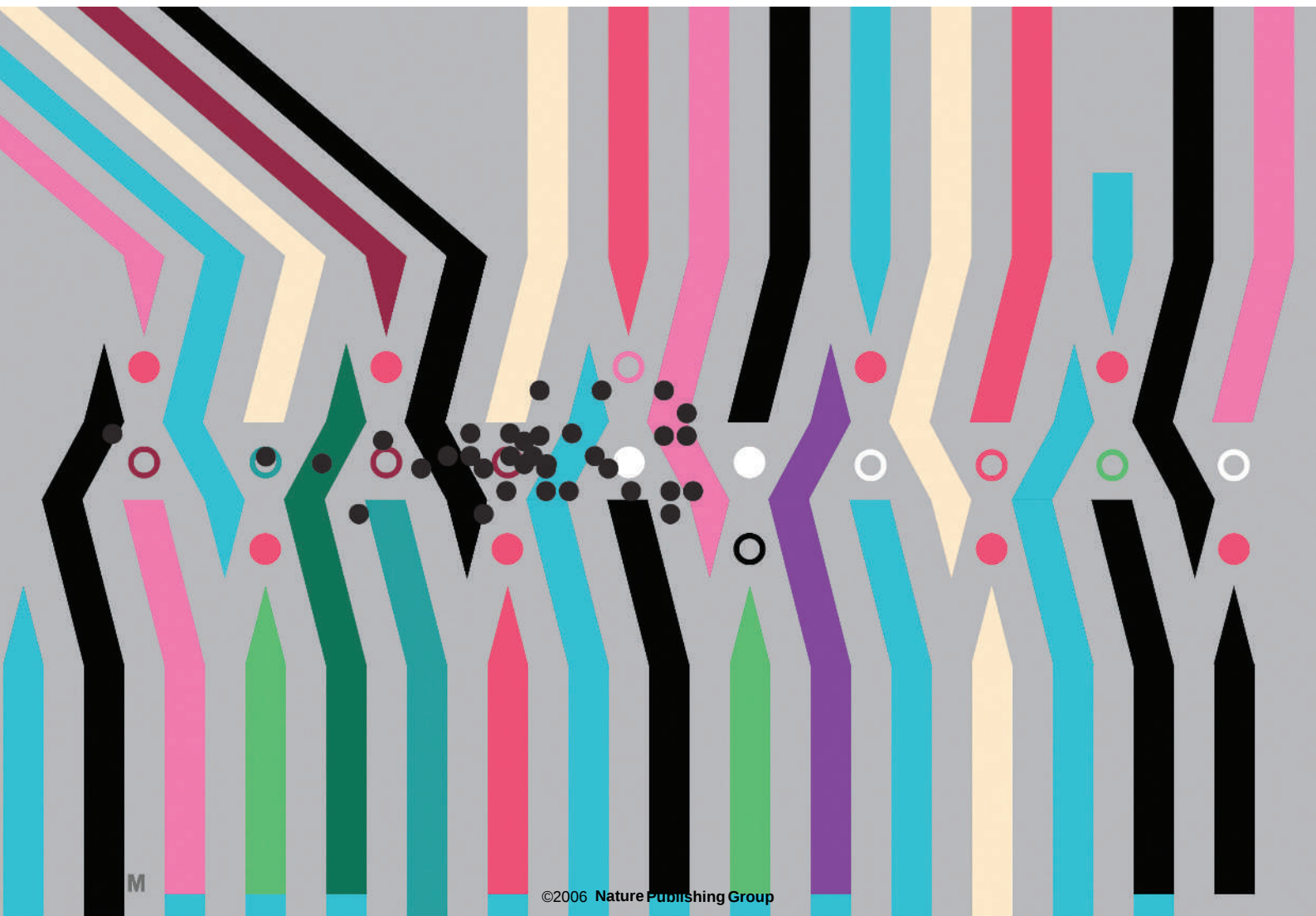
Nevertheless, as a scientific research tool the quantum computer could be revolutionary because of its ability to simulate other quan-

tum systems. In conventional, or classical, computers, information is stored as strings of bits: binary digits each of which can take the value of 0 or 1. The same is true for quantum computers, except that this time the binary digits — 'qubits' — are stored in the quantum states of microscopic systems, such as the electronic state of an atom or ion. So by its very nature, a quantum machine should be much better suited to simulating quantum systems than a classical computer.

A quantum simulator would describe and predict the structure and reactivity of molecules and materials by accurately capturing their fundamental quantum nature. This is the sort of employment the early machines are likely to find: doing calculations of interest to chemists, materials scientists and possibly molecular biologists, says Steane.

"Just a few dozen qubits may shed light on

J. MAGEE



M. BARRETT & J. JOST

other physics problems that are intractable with conventional computers," notes Monroe. "There are models of high-temperature superconductivity and other condensed-matter systems that might be approached in such a quantum simulator."

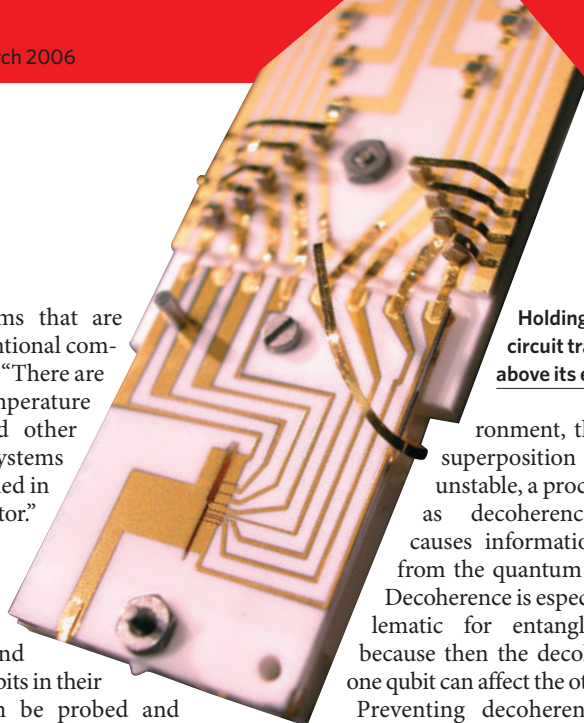
In a spin

In fact, quantum simulations can already be done using atoms and molecules that store qubits in their nuclear spin and can be probed and manipulated using nuclear magnetic resonance (NMR) techniques. In their own terms, these 'computers' "run rings around any classical supercomputer", says Seth Lloyd, a theorist at MIT. He and his MIT colleague David Cory have been using this technique to simulate a variety of quantum systems in crystals of calcium fluoride and other materials. "As the crystal contains a billion billion spins, these simulations remain out of the reach of the most powerful classical computers," says Lloyd. The approach remains limited in terms of the different systems it can simulate, although Lloyd anticipates that fully functioning simulators will be readily available by 2020.

The key to the potential success of quantum computers is also the cause of the problems within the field: the quantum nature of data storage and manipulation. In classical computers, bits have clearly defined values of 1 or 0, but the laws of quantum mechanics allow qubits to exist in a 'superposition' of states — a mixture of both 1 and 0 that would be impossible in an everyday computer. This means that a quantum computer has much greater capacity for storing information.

A quantum processor can also compute with more than one qubit at once by exploiting another quantum property called entanglement, which makes qubits interdependent. The weird nature of the entangled state means that a measurement on one qubit instantly affects another, even though their previous individual states were undefined until that moment. Entangled states don't readily exist in nature: quantum engineers have to make them by allowing qubits to interact with one another.

By exploiting superpositions, a single quantum computer in effect mimics a whole suite of classical computers running at once, and by using entanglement these 'parallel computers' can be linked together. Unfortunately, this powerful parallel processor has an Achilles' heel. A quantum superposition has to remain stable for at least as long as it takes to do the computation. But as soon as qubits interact with their envi-



Holding pen: this circuit traps ions above its electrodes.

ronment, the delicate superposition becomes unstable, a process known as decoherence, which causes information to leak from the quantum computer. Decoherence is especially problematic for entangled states, because then the decoherence of one qubit can affect the others too.

Preventing decoherence means reducing uncontrolled interactions with the environment. Cooling the quantum system to very low temperatures helps — but it may also be necessary to shield the qubits from stray electromagnetic fields. In practice, researchers have found it difficult to avoid decoherence of specific qubits for longer than a few seconds. But in principle it should be possible. "For qubits encoded in trapped ions, nobody really believes that we will ever be limited by coherence time," says Monroe.

Despite the fact that qubits need to be isolated from their environment to avoid decoherence, they must interact strongly with one another, to perform computations. And it must be possible for qubits in superposition to interact strongly with the environment when needed, so that the information can be read out. It is an extraordinarily delicate balancing act, which involves rules that defy intuition and aren't even completely understood.

An easy mistake to make

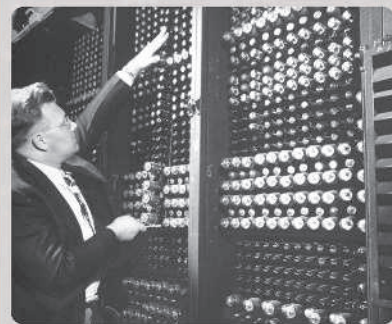
Decoherence also means that, as they process qubits using logic gates, quantum computers will inevitably incur errors at a much higher rate than classical computers. "The modern transistor has an error rate of less than 1 in 10^{14} or more switching events. In comparison, the best quantum gates we currently imagine will optimistically have an error rate of something like 1 in 10^7 ," says Chuang. Some researchers thought at first that this would make quantum computers too error-prone to be useful. But thanks to quantum error-correcting codes devised in the 1990s^{1,2}, it is now possible to correct error rates as high as 1 in 10^5 .

By 2002 the key principles behind a quantum computer had been sketched out by theorists (see 'How to build a quantum computer', overleaf), but how best to implement them in a real device remains a wide-open question. Much of the current effort is focused on making quantum computers using atoms or ions that are held in a trap. In an ion-trap computer, the qubits are encoded in the electronic states

MILESTONES IN SCIENTIFIC COMPUTING

PRE 1960s>>

1946 ENIAC, widely thought of as the first electronic digital computer, is formally unveiled. Designed to compute ballistics during the Second World War, it performs calculations in a variety of scientific fields including random-number studies, wind-tunnel design and weather prediction. Its first 24-hour forecast takes about 24 hours to do.



1951 Marvin Minsky, later of the Massachusetts Institute of Technology (MIT), builds SNARC, the first machine to mimic a network of neurons.

1954 John Backus and his team at IBM begin developing the scientific programming language Fortran.

1956 Building on earlier experiments at the University of Manchester, UK, and elsewhere, MANIAC at the Los Alamos National Laboratory in New Mexico becomes the first computer to play a full game of chess. In 1996, IBM's Deep Blue computer will defeat world chess champion Garry Kasparov.



1959 John Kendrew of the University of Cambridge, UK, uses computers to build an atomic model of myoglobin using crystallography data.

>> PROTOTYPES...

>>

of ions that are confined by an electromagnetic field. The ions interact with each other through electrostatic repulsion, and can be entangled by using laser beams to make them jiggle in unison. The quantum states of the ions can be read out by using other lasers to excite fluorescence, the wavelength of which depends on the ion's electronic state.

But the more qubits there are, the harder it is to read out their complex, collective vibrational states. One way to get round this is to hold most of the ions in a reservoir, and to perform each computational step using just a few of them, transferred from the reservoir to a processing chamber. This means that the ions have to be shuttled around without their quantum states being affected, so that they don't 'lose their memory' on the journey.

Charging ahead

Solving these problems is not easy, but recent progress has been encouraging. "Ion-trap chips look well placed to create useful computers before other methods," says Steane. Last December, for example, Monroe's team reported an ion trap built on a semiconductor chip using standard microfabrication techniques³. The trap held individual cadmium ions for more than an hour, while the researchers were able to move the ions smoothly between trapping sites.

David Wineland's group at the National Institute of Standards and Technology in Boulder, Colorado, is pursuing a similar idea in which the ions are trapped above electrodes etched into a chip's surface⁴. "Both methods have the advantage of using established fabrication techniques," says Wineland. "In the end,

How to build a quantum computer

The current US roadmap for the next decade of quantum computing (<http://qist.lanl.gov>) lists several requirements for a working machine⁸:

1. It must be scalable: it needs a set of qubits that can be added to indefinitely.
2. It must be possible to set all of the qubits to a simple initial state, such as all 0.
3. The interactions between qubits must be controllable enough to make quantum logic gates.
4. To perform operations using these gates, the decoherence times must be much longer than the gate-operation time (typically milliseconds to seconds).
5. There must be some readout capability.
6. To 'wire up' the computer's circuitry, it must be possible to convert memory qubits into processing qubits, and vice versa.
7. It must be possible to move processing qubits accurately between specified locations.

the winner might be determined simply by what is easier to fabricate."

Despite his success so far, Monroe is cautious about the long-term prospects. "Many groups are racing to build complex ion-trap chips," he says. "But it's less clear how trapped ions will ultimately compare with other quantum technologies."

Instead of ions, some researchers are encoding qubits using trapped neutral atoms. Atoms have the advantage that they interact more weakly with their environment than ions, but they also interact more weakly with each other.

They can be trapped by laser beams — and by exploiting the interference pattern generated between crossed laser beams, hundreds of atoms can be held within an 'optical lattice', rather like an egg box. To make the atoms interact, the dimples in the egg box can be shifted closer together by adjusting the trapping beams.

One way to perform quantum computing with atoms is to create discrete clusters of entangled atoms in a larger lattice. This was first suggested in the late 1990s by Hans Briegel, Ignacio Cirac, Peter Zoller and their colleagues at the University of Innsbruck in Austria. It is an approach to quantum computations that Deutsch describes as "far easier to implement physically" than other methods for handling qubits.

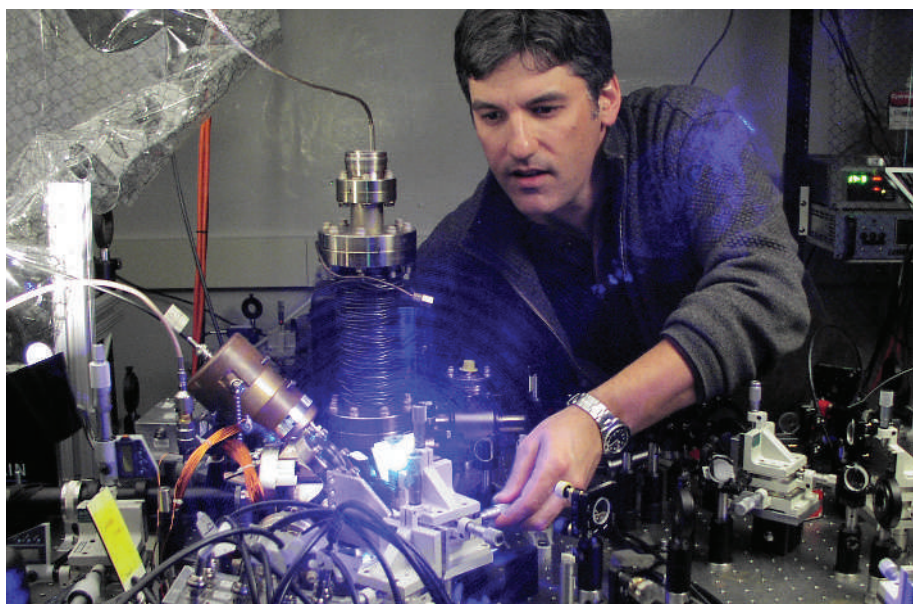
Unlike the standard approach, cluster computation does not involve manipulating individual particles. Instead, before the computation is run, several qubits are brought together in a many-particle entangled state. The answer is then read out at the same time as the computation is actually performed, by making a series of measurements on each individual qubit in the cluster. Its 'one-step' nature makes this an appealing approach, but Cirac, who is now based at the Max Planck Institute for Quantum Optics in Garching, Germany, admits that it requires many more qubits than other methods, and that the error-correction procedures are more elaborate.

Join the dots

There is no shortage of other ideas for building a quantum computer. Some are based on superconducting devices, exploiting the fact that superconductivity is itself a quantum phenomenon. Unlike the systems based on individual particles, the qubits here are superconducting circuits, which hold many-particle quantum states of electrical charge or magnetic flux and can interact through classical electromagnetic forces.

Others hope to create optical quantum computers, encoding qubits into the quantum states of photons, or to make qubits from tiny specks of semiconducting material called quantum dots. "I've been particularly impressed by the advances made in quantum-dot systems and by the superconductor-based approaches," says Chuang. Compared with ion-traps, he explains, they may scale up more easily and are perhaps easier to connect to traditional telecommunication systems for readout.

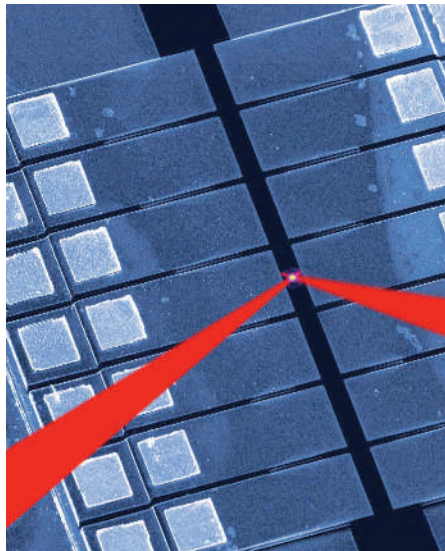
Quantum-dot systems may not produce the first useful computer, says Steane, but they have a naturally faster timescale — largely because the qubits are encoded in electrons, which are much lighter than ions — and so ultimately should outperform other systems such as ion-



Light touch: Chris Monroe aligns laser beams ready to trap ions for a quantum computation.

H.F. MONROE

D. L. STICK



Caught in a trap: a single cadmium ion is held between two electrodes on a semiconductor chip.

trap chips. But, he adds, “the area is still sufficiently open that it is premature to slow efforts on any of the major contenders”.

Because the hurdles in building the hardware are substantial, it is often suggested that the obstacles to making a quantum computer come down to engineering. But there is a bottleneck in theory too. So far, remarkably few specific computational problems have been translated into a form that a quantum machine could run and solve.

In fact, quantum computers are currently little more than two-trick wonders. In 1994, Peter Shor, now based at MIT, devised an algorithm that would allow quantum computers to factor numbers exponentially faster than conventional computers. Factorization is important in cryptography, where it is needed to make and break keys. And in 1996, Lov Grover, who is now at Lucent Technologies in Murray Hill, New Jersey, unveiled a quantum algorithm that can greatly speed up database searches.

Both of these algorithms have already been run using the NMR and optical techniques, but these methods are hard to scale up. At the end of last year, Monroe’s group reported success with a Grover-type search using two cadmium ions in a trap⁵. Admittedly this meant looking through a database of just four entries — hardly a demanding task — but Monroe says the group plans to scale up to dozens of qubits over the next few years.

“It is striking that ten years have passed

since Shor’s invention, and very few new quantum algorithms have been developed,” says Chuang. Among those that have appeared are methods for solving problems in number theory, drawn up by Sean Hallgren at NEC Laboratories in Princeton, New Jersey⁶.

A major stumbling block for those trying to dream up new algorithms is that they first have to identify which problems will benefit most from quantum-computing methods. Theorist Michael Nielsen and his colleagues at the University of Queensland in Australia have recently made progress in this direction by showing that the general problem of finding quantum algorithms can be made easier by borrowing ideas from geometry⁷.

In essence, the number of quantum operations, and thus the length of time, it takes to run an algorithm can be calculated by finding the shortest path between two points in a geometric space defined by all the possible sequences of quantum operations. “It really is a cool idea that has no classical analogue,” says Lloyd. “It opens up a variety of methods for potentially creating new algorithms and for optimizing existing algorithms.”

But not all quantum information processors will need complex algorithms. Many will be purpose-built tools that exploit quantum rules to improve on existing technologies such as atomic clocks and photonic technology. “We’ll probably see rudimentary devices

such as a ‘quantum repeater’ that converts photonic qubits to atomic qubits for error correction, and then back to photons to send them on their way down a long length of optical fibre,” Monroe says.

If that seems a far cry from the quantum brains that are sometimes paraded as the next big thing, we may just have to get used to it. But Lloyd remains upbeat about the prospects. “I agree that quantum computers tailored for specific applications are likely to be built before general-purpose devices. But that doesn’t rule out the possibility that we’ll all be playing quantum *Grand Theft Auto* in the near future.” ■

Philip Ball is a consultant editor for Nature.

1. Shor, P. W. *Phys. Rev. A* **52**, R2493–R2496 (1995).
2. Steane, A. M. *Phys. Rev. Lett.* **77**, 793–797 (1996).
3. Stick, D. et al. *Nature Phys.* **2**, 36–39 (2006).
4. Seidelin, S. et al. preprint at <http://www.arxiv.org/quant-ph/0601173> (2006).
5. Brickman, K.-A. et al. *Phys. Rev. A* **72**, 050306(R) (2005).
6. Hallgren, S. in *Proc. 34th Annu. ACM Symp. Theor. Comput.* 653–658 (ACM, New York, 2002).
7. Nielsen, M. A., Dowling, M. R., Gu, M. & Doherty, A. C. *Science* **311**, 1133–1135 (2006).
8. DiVincenzo, D. P. *Fortschr. Phys.* **48**, 771–783 (2000).

1960s>>

1962 Charles Molnar and Wesley Clark at MIT’s Lincoln Laboratory design the Laboratory Instrument Computer (LINC) for researchers at the National Institutes of Health. It is the first lab-based computer to process data in real time.



1963 In California, the Rancho Arm becomes the first artificial robot arm to be controlled by a computer.

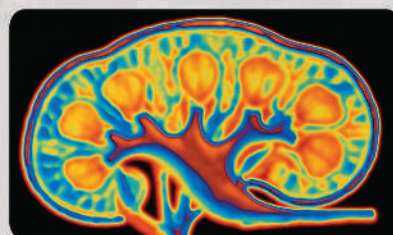
1966 Cyrus Levinthal at MIT designs the first program to represent and interpret protein structures.

1967 ARPANET — the predecessor of the Internet — is proposed by the US Department of Defense for research networking.

1969 Results of the first coupled ocean-atmosphere general circulation model are published by Syukuro Manabe and Kirk Bryan, paving the way for later climate simulations that become a powerful tool in research on global warming.

1970s>>

1971 Computing power shows its potential in medical imagery with a prototype of the first computerized tomography (CT) scanner.



1971 The Protein Data Bank is established at Brookhaven National Laboratory in Upton, New York.

1972 Hewlett Packard releases the HP-35, the first hand-held scientific calculator, rendering the engineer’s slide rule obsolete.

>>

>> MAINFRAMES...

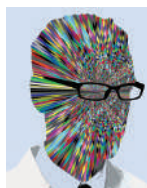
>> WORKSTATIONS...



J. MAGEE

EVERYTHING, EVERYWHERE

Tiny computers that constantly monitor ecosystems, buildings and even human bodies could turn science on its head. **Declan Butler** investigates.



“What are you doing in the lab? Why aren't you out working in the field?” These are not the sorts of question you usually put to your computer. But they should be, according to the proponents of a new type of information technology sometimes known as ‘smart dust’.

In their current, mostly desktop, incarnation, computers used for science usually come into their own quite late in the process of inquiry. Questions are asked, the data that might answer them identified, that data gathered — and only then does the computer start to play a role. In the future, this set up could be reversed. Computers could go from being back-office number-crunchers to field operatives. Twenty-four hours a day, year-in, year-out, they could measure every conceivable variable of an ecosystem or a human body, at

whatever scale might be appropriate, from the nanometric to the continental.

These new computers would take the form of networks of sensors with data-processing and transmission facilities built in. Millions or billions of tiny computers — called ‘motes’, ‘nodes’ or ‘pods’ — would be embedded into the fabric of the real world. They would act in concert, sharing the data that each of them gathers so as to process them into meaningful digital representations of the world. Researchers could tap into these ‘sensor webs’ to ask new questions or test hypotheses. Even when the scientists were busy elsewhere, the webs would go on analysing events autonomously, modifying

their behaviour to suit their changing experience of the world.

If this scenario sounds far fetched, imagine the owner of a mainframe in the 1970s asking why it wasn't sitting on millions of desks and laps worldwide. An absurd question — to which the answer was “it's just a matter of time”. The world's stock of computing power, and the number of devices over which it is distributed, has increased exponentially since then, as has the capacity of networking technology. These trends show no sign of slowing down, and that makes pervasive sensor nets not so much possible as inevitable. One does not need to be a visionary to see that soon, tiny devices with the power of today's desktops will be cheap enough to put everywhere.

Gaetano Borriello, a computer scientist at the University of Washington in Seattle, argues that such widely distributed computing power will trigger a paradigm shift as great as that

“We will be getting real-time data from the physical world for the first time on a large scale.”

— Gaetano Borriello

brought about by the development of experimental science itself. "We will be getting real-time data from the physical world for the first time on a large scale."

Instead of painstakingly collecting their own data, researchers will be able to mine up-to-the-minute databases on every aspect of the environment — the understanding of diseases, and the efficacy of treatments will be dissected by ceaselessly monitoring huge clinical populations. "It will be a very different way of thinking, sifting through the data to find patterns," says Borriello, who works on integrating medical sensors — such as continuous monitors of heart rate and blood oxygen — with their surroundings. "There will be a much more rapid cycle of hypothesis generation and testing than we have now."

Mallikarjun Shankar, who works on sensor webs for military and homeland security at Oak Ridge National Laboratory in Tennessee, agrees. "If one looks at the landscape of computing, this is where it will link with the physical world — where computing science hits the tangible, the palpable world. It is the next frontier."

From virtual to actual

Things don't get much more palpable than the experience of having an offshoot of Europe's biggest ice sheet grinding you into the granite below. That is the lot in life of the sensor web that Kirk Martinez, of the University of Southampton, UK, has been running for the past few years. He is helping glaciologists to study the dynamics of the Briksdalsbreen glacier in northwest Norway, part of the Jostedalbreen ice field, in the hope of better understanding the impact of climate change and weather patterns on the ice field¹.

Martinez's team uses a hot-water 'drill' to place pods — a dozen at any one time — at different locations deep under the ice. Each pod is equipped with sensors that measure variables such as temperature, pressure, and movement; the data collected are used to work out the rate of flow of the glacier and to model subglacial dynamics. The sensor web can be programmed in such a way that the individual pods cooperate. "You can get the pods talking to each other, and deciding that nothing much has happened recently as most of our readings have been the same, so lets the rest of us go to sleep and save batteries, with one waking us up if something starts happening," says Martinez.

"This is where where computing science hits the tangible, the palpable world. It is the next frontier."

— Mallikarjun Shankar

Martinez himself is a computer scientist, not a glaciologist. He was drawn to the task of remotely monitoring a hostile environment around the clock because he wanted the challenge of trying to bring together the various different technologies such sensor webs need. "This is very, very technologically tricky stuff," he says.

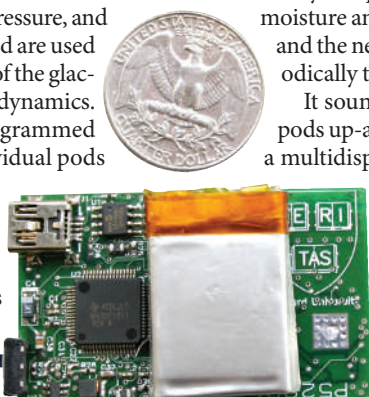
For non-computer scientists, it is even trickier. Researchers can already buy pods the size of cigarette packets or credit cards off the shelf from a slew of new companies such as CrossBow and Dust Networks (both based in the Bay Area of California). But that's only the start. At present, creating a sensor web for a specific scientific application requires extensive customization, particularly in the programming.

Katalin Szlavetz, an ecologist at Johns Hopkins University, in Baltimore, Maryland, works on soil biodiversity and nutrient cycling. She was driven to sensor webs by frustration caused by trying to solve complex problems with limited data. Soil is the most complex layer in land ecosystems, but it is poorly understood, because sampling is limited by the fact that technicians must collect soil samples by hand, and analyse them back in the laboratory. "Wireless sensor networks could revolutionize soil ecology by providing measurements at temporal and spatial granularities previously impossible," she says.

Data stream

Last September, Szlavetz deployed ten motes along the edge of a little stream just outside the university campus. Each mote measures soil moisture and temperature every minute, and the network transmits its data periodically to a computer in her office.

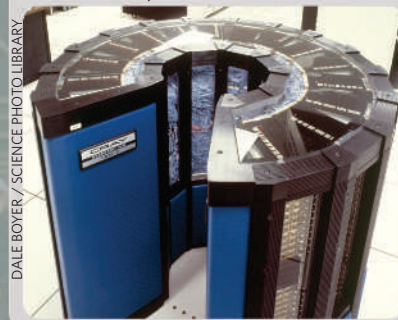
It sounds simple, but just to get the pods up-and-running she had to create a multidisciplinary team, involving computer scientists, physicists and a small army of student programmers. Szlavetz says that she has "no doubt" that pod networks will take off widely, but that it won't happen until they are easier to deploy. And cheaper: "Each unit costs around US\$300, but if you include all the hours of student time, each



Watchful eyes: tiny computers called motes may one day be in everything.

>>

1976 At Los Alamos, Seymour Cray installs the first Cray supercomputer, which can process large amounts of data at fast speeds.



DALE BOYER / SCIENCE PHOTO LIBRARY

1980s>>

1983 Danny Hillis develops the Connection Machine, the first supercomputer to feature parallel processing. It is used for artificial intelligence and fluid-flow simulations.

1985 After receiving reports of a lack of high-end computing resources for academics, the US National Science Foundation establishes five national supercomputing centres.

1989 Tim Berners-Lee of the particle-physics laboratory CERN in Geneva develops the World Wide Web — to help physicists around the globe to collaborate on research.



MICHAEL QUAN/ZUMA PRESS

1990s>>

1990 The widely used bioinformatics program Basic Local Alignment Search Tool (BLAST) is developed, enabling quick database searches for specific sequences of amino acids or base pairs.

1996 George Woltman combines disparate databases and launches the Great Internet Mersenne Prime Search. It has found nine of the largest known Mersenne prime numbers (of the form $2^n - 1$), including one that is 9,152,052 digits long.

>>

>> PERSONAL COMPUTERS...

>> INTERNET...



KNUDSENS

Science of the future: researchers can keep a constant eye on the flow of a Norwegian glacier by tracking miniature sensors buried beneath the ice.

works out closer to \$1,000.” Despite this, Szlavecz says, the Microsoft-funded pilot was a success, revealing previously unobserved variations in soil microclimate, and showing how rain affects wetting and drying cycles⁷. (Szlavecz’s colleague Alexander Szalay is an author of the commentary on page 413).

Handful of dust

Difficulties like Szlavecz’s are all too common, says Kris Pister, founder and chief technology officer of Dust Networks. It was Pister who, at the University of California, Berkeley, coined the term smart dust to describe his vision of sensors smaller than the eye could see joined into networks larger than the mind could comprehend. Pister built prototype sensor webs with funding from the Defense Advanced Projects Research Agency, which is interested in the technology’s military applications. But he says it was the desire to create more usable systems that led him to get the commercial backing to create Dust Networks. “It was very frustrating that while we could do spectacular demos, we couldn’t give scientists something off-the-shelf, to put in a tree or a river. What people want is the ability to just put sensors out in the environment and get data back.”

He likens the current stage of sensor-web development to the early days of computing. “There are a group of experts at the cutting edge of sensor webs, who have the time and expertise to go in and learn how to use the tools

and do all the neat stuff,” he says. “But for people who are not experts it has been difficult to get in and use it.” He predicts, however, that just as the first web browser, Mosaic, and its successor, Netscape, sparked mass take-up of the World Wide Web, so future, more user-friendly sensor-web tools will generate interest.

Although Pister is interested in scientific applications, his key target is the lucrative industrial market for control systems. He has been contracted by the US Department of Energy to help build ‘intelligent’ self-monitoring lighting systems for factories, offices and homes; the 600 billion kilowatt-hours of lighting used for this purpose account for 30% of total building electricity use. “The next step is about really getting some standards and commercial adoption,” he says. “That will drive cost down and performance up, and then scientific uses will take off.”

Even with today’s sensor-web technology, applications for research are proliferating. The Jet Propulsion Laboratory (JPL) in Pasadena, California, which is responsible for most of NASA’s planetary science, has been running

nine large sensor webs to study among other things, flooding, pollution and microclimates, in settings ranging from botanical gardens to the Sevilleta National Wildlife Refuge in central New Mexico. This month, Kevin Dellin, the head of the JPL sensor-web programme, spun it off into a company, Sensorware Systems.

They mote be giants

But sensor webs currently have major limitations for people doing science in the field, says Deborah Estrin, director of the Center for Embedded Networked Sensing in Los Angeles, California, which operates a suite of land- and sea-based monitoring projects in collaboration with university groups. Estrin says that sensor webs alone are often not sufficient for all monitoring needs, and that the cost of sensors prohibits researchers from obtaining the pod densities often needed for detailed field experiments.

Estrin sees the sensor-web revolution as an important thread in a grander tapestry of global monitoring, which involves billions of dollars being poured into projects to monitor the continents and oceans. The US National Science Foundation’s Ocean Observatories Initiative (OOI), for example, plans to spend \$300 million over the next six years on gigabit ‘backbones’ — fibre-optic carriers of data — across the floor of the Pacific Ocean. On land, the planned National Ecological Observatory Network would enable research on terrestrial

“We could do spectacular demonstrations but we couldn’t give scientists something off-the-shelf, to put in a tree or a river.”

— Kris Pister

ecosystems at regional to continental scales in real time. And underneath the surface, the EarthScope project would explore the four-dimensional structure of the North American continent from crust to core. Integrating local sensor webs and all these other networks is one of the biggest challenges facing the development of observational science, Estrin says.

Such mega-observatories may seem very different from Szlavetz's handful of little sensors alongside a stream in Baltimore. But the principles behind them are strikingly similar: to suck the best real-time data out of the environment as possible. "Instead of handling individual data files you will be handling continual streams of data", says Robert Detrick, chair of the OOI's executive steering committee. "You will be combining inputs from different sensors interactively to construct virtual observatories: sensors will be in everything."

The OOI's aim is to get around the fact that oceanographers tend to see only what their research vessels happen to be traversing at any given time. Checking in on the ocean fibre-optic backbone will be swarms of tiny autonomous submarines. These will carry sensors, go off on sampling missions and return to the backbone to recharge their batteries and upload data. "They can all communicate with one another acoustically," Detrick enthuses. "One can say, 'Hey, I've found an anomaly over here, so come on over'". Static sensors will monitor tectonic plates continuously along their entire length. Episodic events such as earthquakes, volcanic eruptions, and instabilities in ocean currents will

be captured in real time, something that is impossible to do using ships.

The existence of such large networks points to some major challenges down the line, says Estrin. Sensor webs will frequently be just single layers in a stack of data-collecting systems. These will extract information at different temporal and spatial scales, from satellite remote-sensing data down to *in situ* measurements.

Managing these stacks will require massive amounts of machine-to-machine communication, so a major challenge is to develop new standards and operating systems that will allow the various networks to understand each other. Sensors and networks of sensors will need to be able to communicate what their data are about, how they captured and calibrated them, who is allowed to see them, and how they should be presented differently to users with different needs. The lack of standards is not an insoluble problem for sensor webs, says Shankar "but it is slowing the field down by several years".

Catching the moment

Despite the difficulties, the use of sensor webs continues to grow. For all the trouble her first ten pods caused her, Szlavetz is upgrading to a network of 200. And experts see no fundamental obstacles to their eventual ubiquity. "We are well on the road to getting there, and I would argue that on many points we are already there," says Borriello. By 2020, says Estrin, researchers using visualization tools like Google Earth will be able to zoom in, not just on an old satellite image, but on "multiple *in situ* observations of the Earth in real time".

Data networks will have gone from being the repositories of science to its starting point. When researchers look back on the days when computers were thought of only as desktops and portables, our world may look as strange to them as their envisaged one does to us. Although we might imagine a science based so much on computing as being distanced from life's nitty gritty, future researchers may look back on today's world as the one that is more abstracted. To them the science practised now may, ironically, look like a sort of virtual reality, constrained by the artificialities of data selection and lab analysis: a science not yet ready to capture the essence of the real world. ■

Declan Butler is a senior reporter for Nature based in Paris.

1. Martinez, K., Riddoch, A., Hart, J. & Ong, R. in *Intelligent Spaces* (eds Stevenson, A. & Wright, S.) Ch. 9, 125–138 (Springer, 2006).
2. Musualoiu-E., R. et al. *Life Under your Feet: A Wireless Soil Ecology Sensor Network* (submitted) The Third IEEE Workshop on Embedded Networked Sensors (2006).



1996 Craig Venter develops the shotgun technique, which uses computers to piece together large fragments of DNA code and hastens the sequencing of the entire human genome.

1998 The first working quantum computers based on nuclear magnetic resonance are developed.

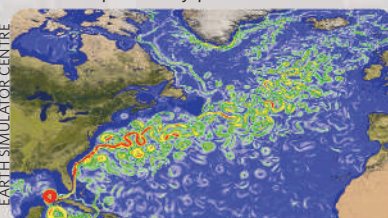
21st CENTURY >>

2001 The National Virtual Observatory project gets under way in the United States, developing methods for mining huge astronomical data sets.



2001 The US National Institutes of Health launches the Biomedical Informatics Research Network (BIRN), a grid of supercomputers designed to let multiple institutions share data.

2002 The Earth Simulator supercomputer comes online in Japan, performing more than 35 trillion calculations each second in its quest to model planetary processes.



2005 The IBM Blue Gene family of computers is expanded to include Blue Brain, an effort to model neural behaviour in the neocortex — the most complex part of the brain.

2007 CERN's Large Hadron Collider in Switzerland, the world's largest particle accelerator, is slated to come online. The flood of data it delivers will demand more processing power than ever before.

Jacqueline Ruttimann



Ecologist Katalin Szlavetz's mini computers tell her about minute-by-minute changes in soil moisture.

BUSINESS

Drive for drugs leads to baby clinical trials

US regulators are moving sharply to ease the early stages of drug development, despite safety concerns. **Meredith Wadman reports.**

David Curiel has just spent four years and US\$4 million building a state-of-the-art plant to make molecules that could deliver gene therapy to patients.

Now it may never be needed. Earlier this year the US Food and Drug Administration (FDA) announced new rules that will allow small doses of experimental drugs to be tested in people before full-scale clinical trials begin. At the same time, it loosened manufacturing requirements for early-stage drug candidates — so Curiel will be free to make his delivery molecules in a cheaper and more convenient workshop.

"We built a palace," says Curiel, director of the gene-therapy centre at the University of Alabama, Birmingham. "If these guidelines had been in place five years ago, we would have saved a lot of time and effort."

Nonetheless, like other medical researchers, Curiel warmly welcomes the rules, which will let him walk down the hall to pick up a batch of vector molecules, instead of crossing eight blocks of the Birmingham campus.

And he won't have to scramble for more money to run the centre: the FDA had told him that it would need five dedicated staff. "Who was going to pay for that?" he asks.

Not everyone is so thrilled with the change. In London last week, six healthy men suffered multiple organ failure after taking part in a phase I trial of a biological drug candidate (see page 388). In light of this, some question whether this is the time to ease up on controls.

Keep the balance

The London event was "absolutely horrific", says Thomas Murray, a bioethicist and president of the Hastings Center in Garrison, New York. "There will be a profound desire not to repeat this experience, and to make sure that the FDA policy changes do not tip the balance too much towards speed and risk."

The rules are part of an FDA drive to open bottlenecks and streamline drug development as estimates of the cost of bringing an original drug to market soar above \$1 billion.

"The problem is that researchers conducting very early studies were required to follow the same manufacturing procedures as companies that mass-produce products for broad-scale distribution," explains Janet Woodcock, the FDA's deputy commissioner for operations.

The changes should benefit academics, she notes, by allowing them to make small batches of experimental drugs and do early testing in people — giving them a better shot at snaring industrial partners to take drugs further.

The manufacturing guidelines relax strict requirements for early-stage drugmaking. For instance, they allow more than one drug to be made at the same facility, and specify that production and quality control can be done by the same person in small operations. Until January, requirements about everything from ventilation to staffing levels had put drugmaking beyond the reach of most university researchers.

"It's a tremendous opportunity for us to rapidly explore many compounds."

The rules on testing in people will mostly benefit the large pharmaceutical companies, by allowing 'phase zero' or 'exploratory' trials. These are brief trials — of seven days or less — in which human subjects are given very low doses of experimental drugs before standard *in vitro* and animal testing is complete.

Phase zero studies do not examine safety or effectiveness; instead, they gather data on the targeting, action and metabolism of a drug in the body. The goal is twofold: to allow drugmakers to identify and jettison failing candidates early, and to generate data that will help them design smarter phase I studies of promising compounds.

That has drawn plaudits from drugmakers, who are frequently forced to use animal data alone to choose the one drug from a panel of candidates that will be propelled into phase I trials.

"We are very interested in pursuing this," says Tim Wright, the deputy head of exploratory clinical development at Swiss drugmaker Novartis. "It's a tremendous opportunity for us to rapidly explore multiple compounds" with limited laboratory and animal testing, he says, adding that his company hopes to

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Fast track: allowing people to take tiny doses of experimental drugs could speed the route to market.

start several of the early trials within a year.

Pfizer has already conducted two phase zero studies, and is planning more, according to Liam Ratcliffe, the company's global chief of clinical research and development. He estimates that one of the studies, conducted on eight volunteers last September, shaved five months off the development time of an anti-coagulant. "It's a welcome change," he says.

Time or money?

The phase zero trials eat up some time before phase I can begin, however, and this may deter small biotechnology companies from doing them. Julian Adams, an organic chemist and chief scientific officer at Infinity Pharmaceuticals, a Massachusetts developer of cancer drugs, says that he has had two opportunities so far to run phase zero trials. He rejected both. "It just wasn't worth it for us," he says. "It saves some money up front. But it doesn't save you time."

At the US National Cancer Institute, however, plans are under way to make use of both changes to the system, says Joe Tomaszewski, deputy director of cancer treatment. He says that production and preclinical toxicology now cost between \$1 million and \$1.5 million for a typical cancer drug. "To get into a phase zero trial, you could cut that in half. So you could put twice as many compounds in the clinic."

Some cancer specialists point out that, at the tiny doses administered in phase zero trials, there isn't going to be any benefit to trial participants. In healthy volunteers, this need

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

GETTY IMAGES

not be an issue. But in cancer, “you’re dealing with dying patients,” says Cy Stein, an oncologist at Albert Einstein College of Medicine in New York. “In phase I, at least we can tell them: ‘We really think we’ve got something here.’” But in phase zero, he says, “you’re offering nothing. I can’t agree to that. Certainly not to save big pharma some time.”

Others, including Sidney Wolfe, director of the health-research arm of Public Citizen, a Washington-based advocacy group, have argued that phase zero studies are ethically troublesome. By reducing the preclinical testing required before an experimental drug goes into humans, says Wolfe, the FDA “increases the risk to the subjects”.

Senator Chuck Grassley (Iowa, Republican) put it more bluntly in a statement when the regulations were released. “The FDA is loosening the reins on drug companies,” he said. “I’m concerned for those who will be receiving these experimental drugs.”

But Woodcock says the approach will protect patients. “Study in people early in the process is going to decrease human exposure to compounds that ultimately fail,” she says, “Which right now is the majority of them.”

Curiel, for his part, hopes to capitalize on the manufacturing changes, cannibalizing the first-class hardware and equipment of his white-elephant facility and moving it to a corner of his existing lab. Complying with the old rules was “complicated, difficult and time consuming,” he says. “The new guidelines will make all of this dramatically easier.” ■

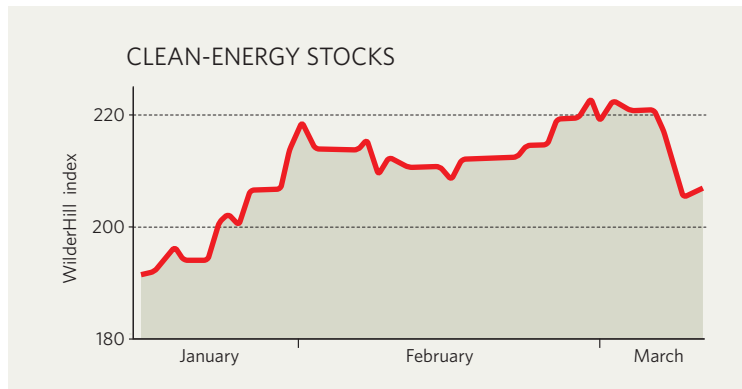
IN BRIEF

UNLUCKY DIP Shares in Human Genome Sciences, the Maryland biotechnology company founded by William Haseltine, tumbled after it announced mixed results from trials of its most promising drug candidate, the hepatitis C treatment *Albuphener*. The company said that the drug’s performance exceeded that of the current market leader, Roche’s *Pegasys*, in patients who took it every two weeks — but not in those who took it once a month. Analysts said the result would lessen the drug’s sales potential, and the company’s shares dropped \$2.80 to \$11 on the news.

FREE RIDER Microsoft has applied to re-register two of its Irish subsidiaries as ‘unlimited liability’ companies — which won’t have to declare any financial results — after coming under fire for allegedly using them as tax shelters. The two companies, Round Island One and Flat Island Company, operate out of small offices in Dublin. They are said to own much of the software company’s intellectual property, and between them reported profits of more than \$4 billion in 2004. The US Treasury is planning to tighten its rules on US companies that hoard their intellectual property in foreign countries that have low corporate tax rates.

CONTRASTING FORTUNES Biochemist Arthur Levinson, the chief executive of Genentech, earned \$69 million last year after cashing in some stock options, the company reported. Three other managers at the thriving California biotechnology company also got more than \$20 million each. But some bosses do it for the love alone: Apple Computer’s chief executive Steve Jobs sat on his options in 2005, and his compensation package for the year was a nominal salary of one dollar — the same as the year before.

MARKET WATCH



January 2006 featured an unexpected endorsement of clean energy by the president of the United States.

But although shares in clean-energy companies reached their highest level since 2001 around the time of George W. Bush’s State of the Union address, they then fell back somewhat.

The WilderHill index, which tracks clean-energy firms that are listed on stock markets in the United States, enjoyed strong growth through 2005 and rose by more than a quarter in the month of January, before slipping back in early March.

Robert Wilder, whose company compiles the index, said that the slip was due partly to shareholders cashing in their profits, and partly to a slump at Distributed Energy Systems, a company that integrates power from different sources. Shares in the Connecticut-

based firm fell by a third on 8 March after it issued a weak earnings forecast.

But Michael Liebrich, head of London-based New Energy Finance and a member of the index’s advisory board, says the fundamentals of the sector remain powerful. “In 2005, there was US\$44 billion of investment in renewable energy worldwide — about 7% of all energy investment,” he says bullishly. “In which direction is that going to go?” He thinks it can grow fivefold in ten years.

And although Wilder sees Bush’s speech as significant, he says that solid financial progress — exemplified by the \$125-million contract that Evergreen Solar of Massachusetts won from a German wholesaler on 15 March — is the source of the sector’s strength.

Colin Macilwain ■

Scientists must be able to report without censorship

SIR — Government scientists must be able to research and report their findings to the public without fear of censorship or intimidation. We need honest results from our science agencies that we can count on. And taxpayers have the right to know the facts.

But in recent weeks, the press has reported allegations that scientists at NASA and the National Oceanic and Atmospheric Administration (NOAA) are routinely silenced when reporting their findings on climate change (“US scientists fight political meddling” *Nature* **439**, 896–897; 2006). If true, this is unacceptable.

That’s why I’ve called for the Government Accountability Office to review the policies and practices of our federal physical-science agencies to ensure openness in communication of their science results. This includes not only NASA and NOAA, but all the federal science agencies within the jurisdiction of the Appropriations Subcommittee on Commerce, Justice, Science, and Related Agencies, on which I serve as the senior Democrat.

US citizens deserve to know what’s happening to their environment from the agencies they rely on to do the research to keep them safe.

Barbara A. Mikulski

United States Senate, 503 Hart Senate Office Buildings, Washington DC 20510, USA

Scientists should be heard, but not expect to set policy

SIR — Your Editorial “Science under attack” (*Nature* **439**, 891; 2006) and News story “US scientists fight political meddling” (*Nature* **439**, 896–897; 2006), on friction between scientists and the Bush administration, impart a curious and unwarrantedly broad meaning to the term ‘unitary executive’ with reference to the US polity. In fact, the term stems from the first sentence of Article II of the US Constitution: “The executive power shall be vested in a President of the United States of America”. Our president has sole (‘unitary’) constitutional power to run the executive branch, and also sole responsibility for all its actions. As Harry Truman put it when he was in office, “The buck stops here.” Unitary executive power does not infringe on the legislative or judicial powers of other branches of government.

The problem that has arisen between the administration and the scientific community is how the notion of presidential executive power is to be extended (or not) to scientific employees of the multitudinous departments

and agencies of the executive branch, and in particular to their strictly scientific opinions (to the extent that those can be separated from policy conclusions). The constitution does not clarify this point, because its authors did not foresee how large the executive branch would eventually become. The issue is neither easy nor trivial, for government scientists have many opinions (some opposed or contradictory), but no independent authority and no responsibility for public policy.

The climate of the present dispute would benefit if both sides would tone down the rhetoric and ease back on the partisan throttle. No administration can gain much by ignoring or silencing the best scientific advice available, either within or outside the government. On the other hand, unelected government scientists in the civil service are not independent political players, and have no inherent licence to make pronouncements designed to call established public policy into question. We govern ourselves primarily through elected officials and their appointees, not by scientific consensus — be it wise or foolish.

William R. Dickinson

Department of Geosciences, University of Arizona, Tucson, Arizona 85718, USA

Fraud: anonymous ‘stars’ would not dazzle reviewers

SIR — In the current discussion about fraud (see Correspondence, *Nature* **439**, 782–784; 2006), the important issue of what actually makes people cheat in the first place has not been addressed. Whether an individual cheats and lies is dependent on many factors, but instant personal advantage is the central one.

Unlike nanotechnology and genetics research, where lucrative patents can beckon, my field of Earth sciences has little to offer but basic research in public institutions. The pressure on the relatively few permanent jobs available to a growing number of scientists in this field is therefore high, and hence the rewards of publishing in the highest-ranked journals are extremely significant for one’s career. The temptation to fumble with the data a bit for the sake of the story, or to include a big-shot author as a showcase, is understandable.

In the most notorious recent cases (the Korean stem-cell work and Jan Hendrik Schön’s nanotechnology work), the peer-review system must be said to have failed, as the frauds were unveiled by people from outside the immediate process. Were the referees the weakest link, and were both the editors and the referees blinded by the aura of the authorships?

Despite some disadvantages, anonymous peer-review remains the fairest way to prevent publication bias (see “Three cheers for peers”

Nature **439**, 118; 2006). But who ensures the quality of the referees and the refereeing process? Clearly, it is the editor’s responsibility to overlook lobbying by authors and referees alike and to improve procedures if necessary to ensure fair assessment of all submissions, not favouring those from ‘star’ authors.

I believe it is best for the publisher not to reveal authors’ names or affiliations to the referee until the submitted manuscript has been accepted or finally rejected. Withholding the authors’ identities in this way will not only ensure high quality, but will stop rejected authors from believing that their submission had been treated unfairly compared with those of ‘star’ authors.

Henning Bauch

Akademie der Wissenschaften und der Literatur, Mainz c/o IFM-GEOMAR, Leibniz-Institut für Meereswissenschaften, Wischhofstrasse 1–3, 24148 Kiel, Germany

Research skewed by stress on highest-impact journals

SIR — Emilio Artacho, in Correspondence (“Reader-appeal should not outweigh merit of research” *Nature* **439**, 534; 2006), raises an important point concerning the growing tension between merit and appeal of research, and alludes to the increasing pressure on young scientists to publish in journals such as *Nature*. This seems to have arisen from senior administrators, both at universities and at funding bodies, requiring simplistic measures of esteem such as numbers of *Nature* papers and citation rates. It creates problems in appointments, tenure and promotions, particularly for sciences that do not traditionally publish in *Nature* or have small communities and thus lower citation rates. As a result, many science departments are now skewed towards research that appeals to the more general reader.

I am pleased that a number of the UK Research Assessment Exercise panels have stated that they will judge papers submitted to them, not on the basis of where they have been published, but rather on their inherent impact and importance. If we want to relieve some of the pressures on young scientists and create a more balanced science community, then we need to re-educate ourselves, and our senior university administrators, so that publishing in journals such as *Nature*, though important, is recognized as just one of a rich variety of research outputs.

Mark Maslin

Department of Geography, University College London, London WC1H 0AP, UK

Contributions to Correspondence may be submitted to corres@nature.com. They should be signed by no more than three authors; preferably by one.

COMMENTARY

Exceeding human limits

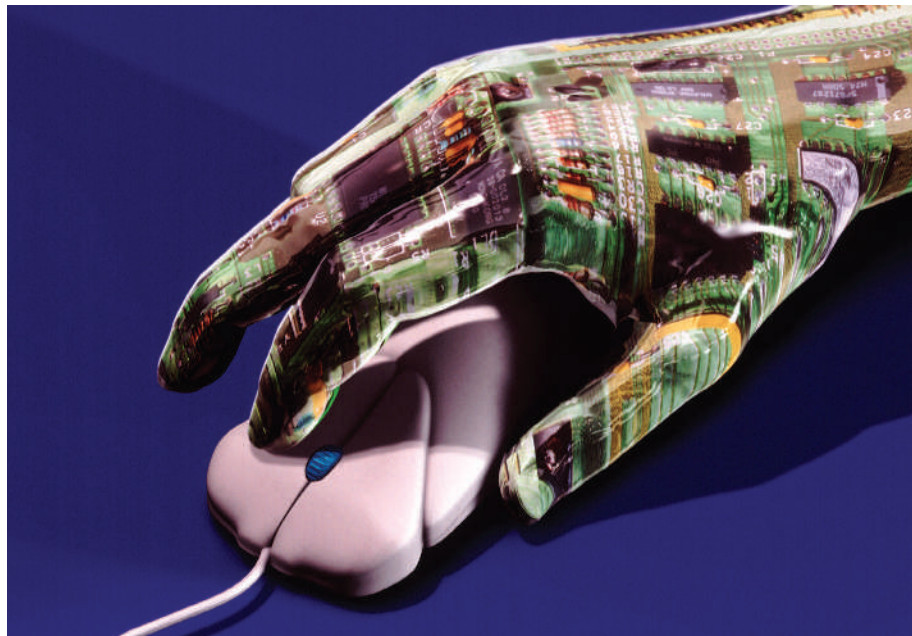
Scientists are turning to automated processes and technologies in a bid to cope with ever higher volumes of data. But automation offers so much more to the future of science than just data handling, says **Stephen H. Muggleton**.



The collection and curation of data throughout the sciences is becoming increasingly automated. For example, a single high-throughput experiment in biology can easily generate more than a gigabyte of data per day, and in astronomy automatic data collection leads to more than a terabyte of data per night. Throughout the sciences the volumes of archived data are increasing exponentially, supported not only by low-cost digital storage but also by the growing efficiency of automated instrumentation. It is clear that the future of science involves the expansion of automation in all its aspects: data collection, storage of information, hypothesis formation and experimentation (see table). Future advances will have the ability to yield powerful new forms of science that could blur the boundaries between theory and experiment. However, to reap the full benefits it is essential that developments in high-speed automation are not introduced at the expense of human understanding and insight.

During the twenty-first century, it is clear that computers will continue to play an increasingly central role in supporting the testing, and even formulation, of scientific hypotheses. This traditionally human activity has already become unsustainable in many sciences without the aid of computers. This is not only because of the scale of the data involved but also because scientists are unable to conceptualize the breadth and depth of the relationships between relevant databases without computational support. The potential benefits to science of such computerization are high — knowledge derived from large-scale scientific data could well pave the way to new technologies, ranging from personalized medicines to methods for dealing with and avoiding climate change¹.

In the 1990s it took the international human genome project a decade to determine the sequence of a single human genome; but pro-



FIREFLY PRODUCTIONS/CORBIS

Many aspects of science are already unsustainable without the aid of computers.

jected increases in the speed of gene sequencing imply that before 2050 it will be feasible to determine the complete genome of every individual human being on Earth. Owing to the scale and rate of data generation, computational models of scientific data now require automatic construction and modification. We are seeing a range of techniques from mathematics, statistics and computer science being used to create scientific models from empirical data in an increasingly automated way. For instance, in meteorology and epidemiology, large-scale empirical data are routinely used to check the predictions of differential-equation models concerning climate variation and the spread of diseases.

Meanwhile, machine-learning techniques from computer science (including neural nets and genetic algorithms) are being used to automate the generation of scientific hypotheses from data. Some of the more advanced forms of machine learning enable new hypotheses, in the form of logical rules and principles, to be extracted relative to predefined background knowledge. This background knowledge is for-

mulated and revised by human scientists, who also judge the new hypotheses and may attempt to refute them experimentally. For example, within the past decade researchers in my group have used inductive logic programming (a subdiscipline of machine learning) to discover key molecular substructures within a class of potential cancer-producing agents². Building on the same techniques, we have more recently been able to generate experimentally testable claims about the toxic properties of hydrazine from experimental data — in this instance, from analyses of metabolites in rat urine following low doses of the toxin³.

Mixing maths

In other sciences, the reliance on computational modelling has arguably moved to a new level. In systems biology, the need to account for complex interactions within cells — in gene transduction, signalling and metabolic pathways — is requiring new and richer systems-level modelling. Traditional reductionist approaches in this area concentrated on understanding the functions of individual genes in isolation. However, genome-wide instrumentation, including microarray technologies, are leading to a system-level

CHANGES TO TRADITIONAL SCIENCE WITH AUTOMATION

Traditional science	Automated science
Hypotheses	Machine-encoded logical hypotheses
Chemical knowledge	Machine-encoded chemical algebra
Experiments	Chemical Turing machine programs
Experimental design	Decision theory

approach to biomolecules and pathways, and to the formulation and testing of models that describe the detailed behaviour of whole cells. This is new territory for the natural sciences and has resulted in multidisciplinary international projects such as the virtual E-Cell⁴.

One obstacle to rapid progress in systems biology is the incompatibility of existing models. Often models that account for the shape and charge distribution of individual molecules need to be integrated with models describing the interdependency of chemical reactions. However, differences in the mathematical underpinnings of, say, differential equations, bayesian networks and logic programs make integrating these various models virtually impossible. Although hybrid models can be built by simply patching two models together, the underlying differences lead to unpredictable and error-prone behaviour when changes are made.

One encouraging development in this respect is the emergence within computer science of new formalisms⁵ that integrate, in a sound fashion, two major branches of mathematics: mathematical logic and probability calculus. Mathematical logic provides a formal foundation for logic programming languages such as Prolog, whereas probability calculus provides the basic axioms of probability for statistical models, such as bayesian networks. The resulting 'probabilistic logic' is a formal language that supports statements of sound inference, such as 'The probability of A being true if B is true is 0.7'. Pure forms of existing probabilistic logic are unfortunately computationally intractable. However, an increasing number of research groups have developed machine-learning techniques that can handle tractable subsets of probabilistic logic⁶. Although it is early days, such research holds the promise of sound integration of scientific models from the statistical and computer-science communities

Miniature roboscientists

Statistical and machine-learning approaches to building and updating scientific models typically use 'open loop' systems with no direct link or feedback to the collection of data. A robot-scientist project in which I was involved offers an important exception⁷. Here, laboratory robots conducted experiments on yeast (*Saccharomyces cerevisiae*) using a process known as 'active learning'. The aim was to determine the function of several gene knockouts by varying the quantities of nutrient provided to the yeast. The robot used a form of inductive logic programming to select experiments that would discriminate between contending hypotheses. Feedback on each experiment was provided by data reporting yeast survival or death. The robot strategy that worked best

(lowest cost for a given accuracy of prediction) not only outperformed two other automated strategies, based on cost and random-experiment selection, but also outperformed humans given the same task.

One exciting development that we might expect in the next ten years is the construction of the first microfluidic robot scientist, which would combine active learning and autonomous experimentation with micro-

fluidic technology. Scientists can already build miniaturized laboratories on a chip using microfluidics⁸ controlled and directed by a com-

puter. Such chips contain miniature reaction chambers, ducts, gates, ionic pumps and reagent stores, and allow for chemical synthesis and testing at high speed. We can imagine miniaturizing our robot-scientist technology in this way, with the overall goal of reducing the experimental cycle time from hours to milliseconds. With microfluidic technology, each chemical reaction not only requires less time to complete, but also requires smaller quantities of input materials, with a higher expected yield. On such timescales it should become easier for scientists to reproduce new experiments and refute their hypotheses.

Today's generation of microfluidic machines is designed to carry out a specific series of chemical reactions, but further flexibility could be added to this tool kit by developing what one might call a 'chemical Turing machine'. The universal Turing machine, devised in 1936 by Alan Turing, was intended to mimic the pencil-and-paper operations of a mathematician. The chemical Turing machine would be a universal processor capable of performing a broad range of chemical operations on both the reagents available to it at the start and those chemicals it later generates. The machine would automatically prepare and test chemical compounds but it would also be programmable, thus allowing much the same flexibility as a real chemist has in the lab.

One can think of a chemical Turing machine as an automaton connected to a conveyor belt containing a series of flasks: the automaton can move the conveyor to obtain distant flasks, and can mix and make tests on local flasks. Just as Turing's original machine later formed the theoretical basis of modern computation, so the programmability of a chemical Turing machine would allow a degree of flexibility far beyond the present robot-scientist experiments, including complex iterative behaviour. In the same way that modern-day Turing machines (computers) are constructed from integrated circuitry, thereby combining the power of many components, a

universal robot scientist would be constructed from a mixture of microfluidic machines and integrated circuitry controllers.

Human touch

This microfluidic Turing machine is not only a good candidate for the next-generation robot scientist, it may also make a good model for simulating cellular metabolism. One can imagine an artificial cell based on a chemical Turing machine being used as an alternative to *in vivo* drug testing. The program running this machine would need to contain algorithms both for controlling the experiment and for conducting the cell simulation. It would represent a fundamental advance in the integration of computation with its environment.

Some may argue that in the context of biological experimentation, the series of chemical reactions is the computation itself. However, one can imagine taking the integration between experiment and environment even further. In particular, by connecting the input and output ducts of the microfluidic Turing machine to the chemical environment of a living cell, one could conduct experiments on cell function. Such levels of close integration between computers, scientific models and experimental materials are, however, still a decade or more away from becoming standard scientific practice.

Despite the potential benefits, there is a severe danger that increases in speed and volume of data generation in science could lead to decreases in comprehension of the results. Academic studies on the development of effective human-computer

"There is a severe danger that increases in speed and volume of data generation could lead to decreases in comprehensibility."

interfaces⁹ emphasize the importance of cognitive compatibility in the form and quantity of information presented to human

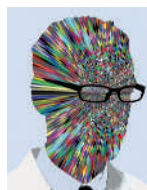
beings. This is particularly critical for technologies associated with hypothesis formation and experimentation. After all, science is an essentially human activity that requires clarity both in the statement of hypotheses and their clear and undeniable refutation through experimentation.

Stephen H. Muggleton is in the Department of Computing and the Centre for Integrative Systems Biology at Imperial College London SW7 2BZ, UK.

1. *Towards 2020 Science* (Microsoft, 2006) <http://research.microsoft.com/towards2020science>
2. Sternberg, M. J. E. & Muggleton, S. H. *QSAR Combinatorial Sci.* **22**, 527–532 (2003).
3. Tamaddoni-Nezhad, A., Kakas, A., Muggleton, S. H. & Pazos, F. in *Proc. 14th Int. Conf. Inductive Logic Programming* 305–322 (Springer, 2004).
4. Takahashi, K. et al. *Bioinformatics* **13**, 1727–1729 (2003).
5. Halpern, J. Y. *Artif. Intell.* **46**, 311–350 (1990).
6. De Raedt, L. & Kersting, K. in *Proc. 15th Int. Conf. Algorithmic Learning Theory* 19–34 (Springer, 2004).
7. King, R. D. et al. *Nature* **427**, 247–252 (2004).
8. Fletcher, P., Haswell, S., Watts, P. & Zhang, X. *Lab-on-a-Chip Micro Reactors for Chemical Synthesis*. Dekker *Encyclopedia of Nanoscience and Nanotechnology* (2001).
9. Jacko, J. A. & Sears, A. *The Human-Computer Interaction Handbook* (Lawrence Erlbaum Associates, 2003).

The creativity machine

What will emerge from using the Internet as a research tool? The answer, Vernor Vinge argues, will be limited only by our imaginations.



We humans have built a creativity machine. It's the sum of three things: a few hundred million computers, a communication system connecting those computers, and some millions of human beings using those computers and communications.

This creativity machine is the Internet. It has already changed the way we do science, most importantly by enhancing collaboration between researchers. The present-day Internet provides convenient connections between computerized labs, simulations and research databases. It also represents an enormous financial investment that is driven by the demands of hundreds of millions of consumers. As such, the total Internet software and infrastructure investment dwarfs the budgets of scientific research programmes and even of many government defence programmes. And more than any megaproject of the past, the essence of the Internet is to provide coordinated processing of information. For researchers seeking resources, these are facts worth considering.

For some disciplines, the Internet itself has become a research tool: grid computing has been used to exploit the power of millions of Internet-connected machines. Building on the popularity of SETI@home — an experiment that uses Internet-connected computers to search for extraterrestrial intelligence — and prime-number hunts, there are now physics, medical and proteomics projects enlisting the enthusiasm of people (and their computers) across the world. For linguists¹ and sociologists, new questions can be investigated simply by observing what occurs on the publicly available Internet. Even experimental sociology is possible: in their study of social influence on music preference, Salganik *et al.*² recruited more than 14,000 subjects through a popular website, ran online trials on these subjects, and then obtained results directly from their experiment website.

The possibilities do not end there. Even online games are attracting academic interest. Some games have millions of players. MMORPGs (massively multiplayer online role-playing games), such as World of Warcraft and EverQuest, feature vivid three-dimensional action involving both cooperation and

combat. Another genre of MMORPGs lack a significant combat or quest element and are more often called 'virtual worlds'. For example, the virtual world Second Life has the visual realism of many MMORPGs, but it exists as a venue for the participants rather than as a pre-designed adventure. Second Life provides a range of software tools, including a programming language, that gives participants the power to create artefacts according to their own designs. Thus the game depends on the skill and creativity of its participants to generate content. Such virtual worlds have already been used for educational projects, and are worthy of psychological and social research.

People power

The notion of enlisting users to create content is widespread on the contemporary Internet. Companies such as Google provide users with tools to integrate search and mapping services into their own websites. Interested users are numerous and have their own resources. In the 1990s, we had an early glimpse of the power of this new creativity machine: computers plus networks plus interested people delivered free and open-source software (FOSS) products of the highest quality, including the GNU/Linux operating system. FOSS products provide low-cost and flexible alternatives to proprietary software. For example, there is at least one open-source virtual-world platform, Croquet³, which allows users to customize and extend its architecture at all levels. FOSS tools can be mixed and matched with proprietary software to deal with an enormous range of projects from quick, *ad hoc* combinations of data harvested from multiple locations⁴ to large, long-duration experiments.

All this points to ways that science might exploit the Internet in the near future. Beyond that, we know that hardware will continue to improve. In 15 years, we are likely to have processing power that is 1,000 times greater than today, and an even larger increase in the

number of network-connected devices (such as tiny sensors and effectors). Among other things, these improvements will add a layer of networking beneath what we have today, to create a world come alive with trillions of tiny devices that know what they are, where they are and how to communicate with their near neighbours, and thus, with anything in the world. Much of the planetary sensing that is part of the scientific enterprise will be implicit in this new digital Gaia. The Internet will have leaked out, to become coincident with Earth.

How can we prepare for such a future? Perhaps that is the most important research project for our creativity machine. We need to exploit the growing sensor/effector layer to make the world itself a real-time database. In the social, human layers of the Internet, we need to devise and experiment with large-scale architectures for collaboration. We need linguists and artificial-intelligence researchers to extend the capabilities of search engines and social networks to produce services that can

bridge barriers created by technical jargon and forge links between unrelated specialties, bringing research groups with complementary problems and solutions together — even when those groups have not noticed the possibility of collaboration. In the end, computers plus networks plus people add up to something

significantly greater than the parts. The ensemble eventually grows beyond human creativity. To become what? We can't know until we get there.

Vernor Vinge is an emeritus professor of computer science at San Diego State University. His novel *Rainbows End* (2006) considers the Internet of 2025.



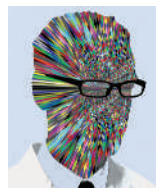
Participants in Second Life use software and creativity to build their environment.

1. Kilgariff, A. & Grefenstette, G. *Comput. Linguist.* **29**, 333–347 (2003).
2. Salganik, M. J., Dodds, P. S. & Watts, D. W. *Science* **311**, 854–856 (2006).
3. Croquet Project www.opencroquet.org
4. Butler, D. *Avian Flu Maps in Google Earth* <http://declanbutler.info/blog/?p=16>

© 2006, LINDEN RES.

Science in an exponential world

The amount of scientific data is doubling every year. **Alexander Szalay** and **Jim Gray** analyse how scientific methods are evolving from paper notebooks to huge online databases.



Scientists are trained early to keep careful records in their laboratory notebooks — recording both experimental procedures and observations, so that they can analyse their results and so that others can replicate what they have done. Galileo did it, Mendel did it, Darwin did it, and we are supposed to do it. This worked fine when small amounts of data were entered into notebooks and the analysis was computed alongside them. But data volumes are doubling every year in most areas of modern science and the analysis is becoming more and more complex, exceeding the capacity of the paper notebook. With data correlated over many dimensions and millions of points, none of the old steps — do experiment, record results, analyse and publish — is straightforward. Many predict dramatic changes to the way science is done, and suspect that few traditional processes will survive in their current form by 2020 (ref. 1).

Today, most scientists have replaced or enhanced their notebooks with desktop computers that record their results, provide a portal to the scientific literature, and link them to collaborators via e-mail. These computers also perform data analysis; Matlab, Mathematica and Excel are popular analysis tools. But none of these programs scale up to handle millions of data records — and they are primitive by most standards. As data volumes grow, it is increasingly arduous to extract knowledge. Scientists must labour to organize, sort and reduce the data, with each analysis step producing smaller data sets that eventually lead to the big picture. Analysing terabytes of data (one terabyte is 1,000 gigabytes) is a challenge; but petabyte data sets (of more than 1,000 terabytes) are on the horizon. One petabyte is equivalent to the text in one billion books, yet many scientific instruments, including the Large Synoptic Survey Telescope, will soon be generating several petabytes annually.

In response to this

data deluge, the systematic use of databases has become an integral part of the scientific process. Databases provide tools to organize large data sets, find objects that match certain criteria, compute statistics about the data, and analyse them to find patterns. Many experiments today load their data into databases before attempting to analyse them. But there are few tools to properly visualize data across multiple scales and data sets. If we can no longer examine all the data on a single piece of paper, how can we 'see' a new pattern or find a data point that does not fit a hypothesis? Fortunately there are database tools, such as data cubes, that we believe can fulfil this role (see 'Data cubes' overleaf).

The same language

Experiments are themselves becoming electronic as computers become essential parts of scientific instruments; they are used not only to manage and analyse vast data sets, but also to acquire them in the first place. Procedures already involve instruments and software with myriad parameters. It is difficult to capture all the model numbers, software revisions, parameter settings and process steps in an enduring format. For example, imagine a measurement taken using a DNA-sequencing machine. The output is cross-correlated with a sequence archive (GenBank) and the results are analysed with Matlab. Fully documenting these steps would be arduous, and there is little chance that

someone could repeat the exact procedure 20 years from now; both Matlab and GenBank will change enormously in that time. As experiments yield more data, and analysis becomes more complex, data become increasingly difficult to document and reproduce.

One might argue that complex biological experiments have always been difficult to reproduce, as there are so many variables. But we believe that with current trends it is nearly impossible to reproduce experiments. We do not have a solution for this problem, but it is important to recognize it as such, and to do what is possible to capture the workflows and to develop protocols for documenting instruments, procedures and measurements in ways that will be usable in several decades' time.

Increasingly, scientists are analysing complex systems that require data to be combined from several groups and even several disciplines. There are collaborations sharing data across departments and time zones, and important discoveries are made by scientists and teams who combine different skill sets — not just biologists, physicists and chemists, but also computer scientists, statisticians and data-visualization experts. It is important to realize that today's graduate students need formal training in areas beyond their central discipline: they need to know some data management, computational concepts and statistical techniques.

A collaboration involving hundreds of Inter-



Automated systems will transform data collections, from astronomy (left) to sampling soil properties under our feet.

DATA CUBES

Traditional notebook and analysis tools are being challenged not just by data volumes, but also by data complexity. For example, the three-dimensional structural representation of a complex protein is not easily transcribed into a notebook.

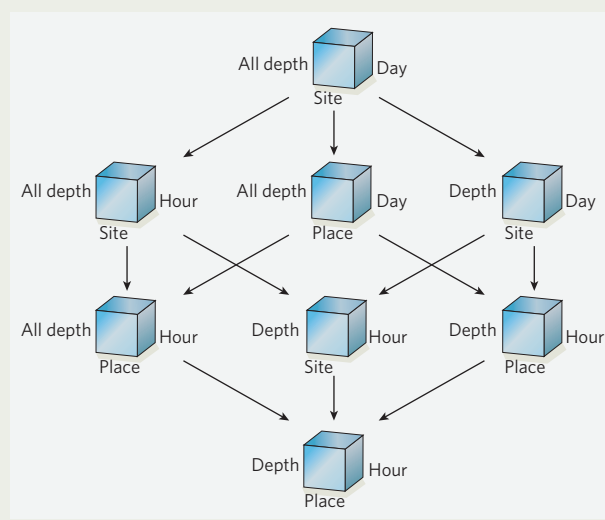
Complex scientific data are often organized as a collection of independent variables and their dependent

measurements.

For example, meteorological data (temperature, pressure, humidity, wind velocity and direction) are collected at various times and locations (latitude, longitude and altitude). These data can be thought of as a multidimensional cube in which time and place exist in four dimensions and the measurements are shown by vectors at each point in the cube

(pictured right).

A meteorologist may ask: show me the minima, maxima and average winds for Australia aggregating over times (hour, day, week, year) and volumes (per square kilometre of the atmosphere). The data cube makes it easy to express such user queries and to compute the answers, even to seemingly complex questions.



C. STOLTE & P. HANRAHAN, STANFORD UNIV./TABLEAU

net-connected scientists raises questions about standards for data sharing. Too much effort is wasted on converting from one proprietary data format to another. Standards are essential at several levels: in formatting, so that data written by one group can be easily read and understood by others; in semantics, so that a term used by one group can be translated (often automatically) by another without its meaning being distorted; and in workflows, so that analysis steps can be executed across the Internet and reproduced by others at a later date.

Standards for sharing data are crucial, for example, in understanding soil ecosystems. We are helping to build a system for measuring long-term environmental trends that affect soil biodiversity (www.lifeunderyourfeet.org; see also News Feature, page 402). This system integrates local environmental data from a sensor network with regional data on hydrology, climate, biodiversity and biogeochemistry. For these data to be useful to others, they must be published using a controlled vocabulary and in standard forms, and the instruments and measurements must be well specified. Fully documenting the sensors and data-collection process is arduous, and there are few standards for us to draw on.

Data gold-mine

Multidisciplinary databases also provide a rich environment for performing science; that is, a scientist may collect new data, combine them with data from other archives, and ultimately deposit the summary data back into a common archive. Many scientists no longer 'do' experiments the old-fashioned way. Instead they 'mine' available databases, looking for new patterns and discoveries, without ever picking up a pipette.

But this data-rich approach to science faces challenges. The speed of the Internet has not kept pace with the growth of scientific data sets. And so large data archives are becoming increasingly 'isolated' in the network sense — one can copy gigabytes across the Internet today, but not petabytes. In the future, working

with large data sets will typically mean sending computations to the data, rather than copying the data to your workstation. But the management of distributed computations raises new questions of security, free access to public data and cost. Few data archives address these issues today.

Are we reaching the limits of what one scientist, or one lab, can expect to achieve in data handling and analysis? If so, this will have implications for how we review and publish our work. For example, a data-mining paper needs to include the explicit description (database query) of how the data that were analysed in the paper were collected and filtered, but not the data themselves. In this way, a reviewer with access to public data could reproduce the data sets and analysis procedures. For the analysis to be repeatable in 20 years' time requires archiving both data and tools.

The publication process itself is increasingly electronic, with new ways to disseminate scientific information (such as the preprint repository arXiv.org). But there is, as yet, no standard for publishing large volumes of data. Paper appendices cannot hold all the data needed to reproduce the results. Some disciplines have created their own data archives, such as GenBank; others just let data show up, and then disappear, on individual scientists' websites. Astronomers created the International Virtual Observatory Alliance (www.IVOA.net), integrating most of the world's medium and large astronomy archives. This required new standards for data exchange, and a semantic dictionary that offers a controlled vocabulary of astronomy terms.

To encourage data sharing, it should be rewarded. Public data creators and publishers should be given credit, and archives must be able to automatically provide provenance details. Current databases have a long way to go to achieve this ideal.

For how long will this exponential growth in scientific data continue? Desktop computers today are as powerful as the super-computers of 10 years ago. Similar progress is happening

with scientific instruments — they quickly become obsolete and are replaced by better and often cheaper ones. Likely computer-performance improvements by 2011 include tenfold more processing, storage and network bandwidth per dollar. So we can expect ten times more data.

Smaller is faster

However, not all experiments will experience exponential growth. There is reason to believe that it will be the smaller experiments, not the big multibillion-dollar facilities, that will grow the fastest. Exponential growth occurs when a new generation of instruments leapfrogs the previous generation, which become obsolete. There are two trends in science today, scaling up and scaling out. Some scientists are building billion-dollar facilities, such as astronomy's Large Synoptic Survey Telescope or the Large Hadron Collider, which are only affordable as international collaborations. Such facilities are not easily leapfrogged. And once these petascale experiments are switched on they will produce roughly the same amount of data each year — merely linear growth. But in the scaling-out model, experiments that deploy an array of small instruments can exploit the coming explosion in cheaper commodity technology. The wireless sensors that were US\$300 a year ago are \$100 today, and will be \$30 next year. A similar phenomenon occurred with DNA chips and gene sequencers. It is important to recognize this pattern; it is universal. And so although some sub-disciplines may reach a plateau in data generation, other technological innovations will take their place. Scientists in 2020 will continue to work in an exponential world. ■

Alexander Szalay is in the Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA.

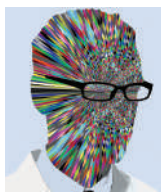
Jim Gray is at Microsoft Research, San Francisco, California 94105, USA.

1. *Towards 2020 Science* (Microsoft, 2006); <http://research.microsoft.com/towards2020science>

Can computers help to explain biology?

The road leading from computer formalisms to explaining biological function will be difficult, but

Roger Brent and Jehoshua Bruck suggest three hopeful paths that could take us closer to this goal.



It doesn't require vast prophetic vision to identify developments in computers and information technology that will greatly affect the practice of biology. By 2020 we expect that biologists will use computers, numerous 'omic' data types (ref. 1) and a greatly expanded biological literature to design experiments, generate and analyse new data, and think about their own work. But we will leave forecasting about PubMed and Google, metadata and the semantic web to others. Instead, we wish to consider some of the formalisms offered by computer science² that developed alongside computing machines. The search for biologically relevant formalisms has a chance to greatly affect the understanding of biological function, in ways we are just starting to imagine.

Today, by contrast with descriptions of the physical world, the understanding of biological systems is most often represented by natural-language stories codified in natural-language papers and textbooks. This level of understanding is adequate for many purposes (including medicine and agriculture) and is being extended by contemporary biologists with great panache. But insofar as biologists wish to attain deeper understanding (for example, to predict the quantitative behaviour of biological systems), they will need to produce biological knowledge and operate on it in ways that natural language does not allow.

Living computers

We begin with what we know in 2006: the trajectory of living systems through developmental time and space is highly determined by the actions and interactions of functional molecules encoded by their genomes. These encoded molecules are further influenced by external perturbations. Because the dynamic behaviour of biological systems is highly determined by a central stored program, living systems differ profoundly from all other naturally occurring, time-evolving systems. The weather has no genome.

Some aspects of biological systems, such as the sequence of encoded proteins (which determines their structure), arise directly from the genome. But others, including most biological functions, arise from the genome by

considerably more complex routes, with the consequence that function typically occurs simultaneously at multiple levels^{3,4}. These levels include the biochemical activity of an individual protein, the function of that protein in cellular processes involving other proteins, and the developmental trajectory of those processes within a multicellular organism^{4,5}. None of these levels is more true or fundamental than the other.

If biology and information science continue with business as usual, then, by 2020, most of the natural-language stories of 2006 about biological function will be subsumed into more sophisticated narratives, which will be better organized and accessed by computers. But the outlines of most of these stories will probably remain unchanged. Here, however, we imagine ways that formalisms from computer science might contribute to a deeper understanding of biological function. These approaches will not bear fruit without deliberate and difficult work.

The development of computer science required both new formalisms to capture reasoning in natural languages and ways to implement those formalisms in physical devices⁶. In 2006, it seems reasonable to compare living systems to 'von Neumann', or stored-program, computers, with processing systems (here encoded by the stored program), various external and internal inputs, and outputs in the form of execution. In this view, the biological system is not primarily a factory or a chemical plant but an assemblage that takes information, processes it, decides and executes.

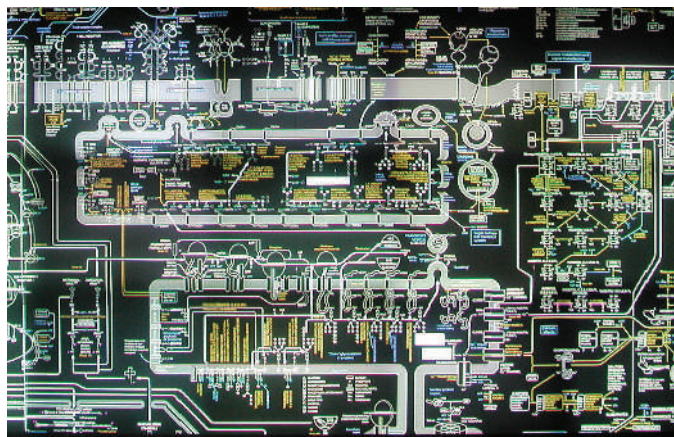
As yet there is no theory that can specify the

meaning or purpose of a string of computer code, but Sussman has suggested how elements of such theory might arise⁷. He points out that mathematics had its roots in a workaday human activity, that of Egyptian surveyors redefining the boundaries of fields after the Nile floods receded. Rigorous thinking about this activity led eventually to mathematics: geometry, trigonometry, algebra and beyond. In the same way, a workaday activity — the design and use of procedural imperative languages to write code ('do this, now, do that, if such a thing happens, then do this!') to program computers — may lead to new formalisms describing information processing and eventually to new mathematics.

Blurred boundaries

In biological systems it seems reasonable to view the DNA script in the genome as executable code, code that could have been specified by a set of commands in a procedural imperative language. And in the same spirit, we can view any signal-transduction pathway as a collection of protein machines that takes inputs from inside and outside the cell, performs processing operations on those inputs to arrive at decisions, and communicates those decisions to an apparatus that executes it.

However, to make the analogy between biological systems and von Neumann computers is to reveal important differences between them (Fig. 1). At the level of cells and organisms, biological systems differ from computers in many ways, including (but not limited to): lack of modularity and boundaries in code; lack of fixed order of execution in code; self-



Biology needs to move beyond natural-language descriptions of biomolecules and pathways. Adopting new formalisms from computing may lead to greater insight than even graphical displays (such as this wall chart) can offer.

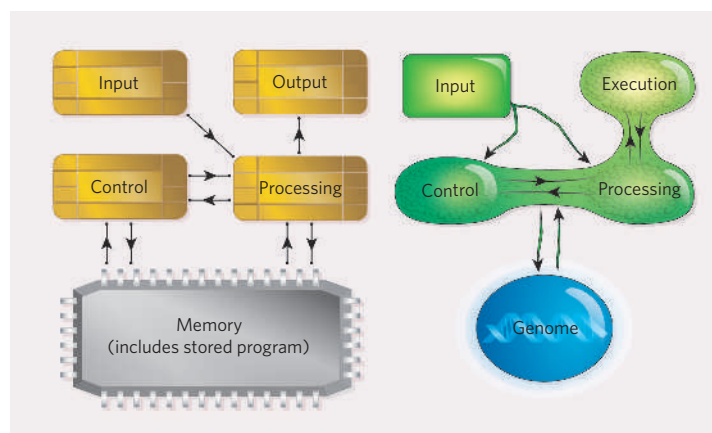


Figure 1 | Biological systems (right) have similarities and differences to von Neumann computers (left). In biological systems there is no distinction between processor and output, as function, phenotype and selection act at many levels.

assembly of encoded components; lack of intelligible sentient design; and lack of crisp boundaries between memory, processor, input and output components.

Most importantly, biological systems usually lack a clear boundary between processing apparatus and output. This distinction arises because function in biology is a consequence of selection, and selection usually acts at many different levels. Thus, the simplest human question ‘what does the system do?’ (which translates into ‘what was the system selected for?’) usually has simultaneous multiple correct answers. This fact will continue to frustrate analyses of biological systems in terms of the ‘objective functions’ they are ‘optimized’ to ‘execute’. Because of this difference, it is unlikely that even a mature theory of stored-program machines will be adequate to explain biological systems. However, even in the absence of grand theory, one can work on intermediate steps. Here we describe three avenues worth exploring.

One fruitful approach formalizes cause-and-effect relationships between named proteins and regulatory sites by translating these into defined chemical reactions undergone by defined molecular species. These reactions can be modelled as differential equations constrained by the rules of chemical kinetics, more formally codified as the ‘chemical master equation’⁸. In biology, differential-equation models have a mixed history; they were vital for understanding transmission of the nerve impulse⁹ and for helping to identify reaction types before the channel molecules were discovered, but were less successful in circadian-rhythm research until biologists identified molecular entities and relevant reactions.

Measure of meaning

For most biological narratives, the resulting sets of differential equations are too complex to be analytically tractable. But their dynamic behaviour can be approached down a second path — by simulating approximate numerical and stochastic methods¹⁰. These simulations already constitute ‘theory’, in the narrow sense that they can generate hypotheses that can then be tested by direct experiment. Equally important, they have inspired mathematicians

and computer scientists to apply existing means to reduce complexity and seek new ones. For example, biological reaction networks do not have an order of execution, but probabilistic methods can be used to explore the most likely chains of reactions executed by a given network (M. Riedel and J. Bruck, personal communication).

A third path to better understanding of function begins with deeper analysis of the natural language now used to describe it. The cause-and-effect stories of function of proteins and regulatory sites use an impoverished vocabulary: many proper nouns, few verbs and some prepositional phrases denoting location. Like information in Geographical Information Systems, which also have a limited vocabulary, biological narratives of cause and effect are readily systematizable by computers. There are at least four commercial companies working to provide such systematizations, which are already providing some insight¹¹.

But for biological function, just beyond cause-and-effect narratives and before the ultimate truths of fitness and selection, there lies a muddy patch of ground known as ‘teleology’. Teleology is hard to avoid: it is difficult to explain why the lens of the eye is transparent without at some point mentioning that the eye is ‘for’ seeing. But in that mud there may be hope.

The twentieth-century architects of information theory⁶ deliberately restricted their concept of information because they were limited by their ability to define and measure it. Information theorists wanted to build a theory that involved the meaning (the semantic content) of messages, but could not measure meaning in sender, recipient or at any point in between. Instead, they chose a meaning for information that was restricted to the carrying capacity of communications channels such as telephone lines — the information technology of their era.

Similarly, biologists would like to cast their descriptions in terms of meaning and purpose, but are limited in their ability to mea-

sure those things. As we have said, biology does offer a clear definition of meaning (‘it was selected’), but the multiple levels at which selection acts means that meaning is always difficult to determine.

Deeper understanding

Happily, there is considerable interest in wanting to build one element of biological semantics — the passage of time — into information theory. Formalizations of information processing that embodied this and other semantic concepts relevant to biology might help biologists to go beyond quantifying reaction rates and molecular species of biological systems to understand their dynamic behaviour. They might also help to suggest new experiments — perhaps on synthetic biological systems engineered to have a crisper division between process and output, which could then be evolved by artificial selection. This approach might bring a deeper understanding of function at its most fundamental level of fitness and selection.

However marvellous developments in computation are by 2020, if their impact is limited to information generation, handling, visualization and integration, it will mean that their potential contribution to a more predictive understanding of biological function will have failed. By laying out three paths from current

“We imagine ways that formalisms from computer science might contribute to a deeper understanding of biological function.”

computer science that might lead to deeper insights, we at least hope to stir things up. But we also observe growing frustration with business as usual. If we knew better how biological systems worked, we

could better perturb existing ones (such as ours, for human medicine) and we could design and build better ones. The fact that both possibilities and frustrations are now starkly evident should make the next 16 years interesting indeed. ■

Roger Brent is at The Molecular Sciences Institute, Berkeley, California 94704, USA; Jehoshua Bruck is at the California Institute of Technology, Pasadena, California 91125, USA.

1. Weinstein, J. N. *Science* **282**, 687 (1998).
2. Simon, H. R. *The Sciences of the Artificial* 1st edn (MIT Press, Cambridge, Massachusetts, 1969).
3. Fields, S. *Nature Genet.* **15**, 325–326 (1997).
4. Kirschner, M. W. *Cell* **121**, 503–504 (2005).
5. Brent, R. *Cell* **100**, 169–183 (2000).
6. Shannon, C. E. & Weaver, W. *The Mathematical Theory of Communication* (Univ. Illinois Press, 1948).
7. Sussman, G. J. *The Legacy of Computer Science NCR/CSTB Symposium on the Fundamentals of Computer Science* (2001).
8. Gillespie, D. T. *Physica A* **188**, 404–425 (1992).
9. Hodgkin, A. L. & Huxley, A. F. *J. Physiol. Lond.* **117**, 500–544 (1952).
10. Gillespie, D. T. *J. Phys. Chem.* **81**, 2340–2361 (1977).
11. Calvano, S. E. *et al. Nature* **437**, 1032–1037 (2005).

Acknowledgements We are grateful to L. Lok, K. Tahashashi, D. Endy, O. Resnekov, P. Rabinow, D. Gillespie, M. Cook and M. Riedel for useful discussions and comments on the manuscript.

A two-way street to science's future

To view the relationship between computing and science as a one-way street is mostly untrue today, argues **Ian Foster**, and will be even less true by 2020.



A growing number of sciences, from atmospheric modelling to genomics, would not exist in their current form if it were not for computers. A simplistic analysis of this relationship focuses on hardware, and sees

science as largely a passive beneficiary of the computing industry's relentless innovation, acquiring and applying to its own ends the fastest computers, largest disks and most capable sequencing machines. In this view, science and computing (as an intellectual discipline) have little to say to each other: it is the computer industry that drives the advances that have an impact on science.

A more sophisticated narrative says that science is increasingly about information: its collection, organization and transformation. And if we view computer science as "the systematic study of algorithmic processes that describe and transform information"¹, then computing underpins science in a far more fundamental way. One can argue, as has George Djorgovski, that "applied computer science is now playing the role which mathematics did from the seventeenth through the twentieth centuries: providing an orderly, formal framework and exploratory apparatus for other sciences."²

Information overload

Why this shift? Of course, there is more information, as technology allows us to collect, store and share vastly more data than before. Equally significant is that science is becoming less reductionist and more integrative, as researchers attempt to study the collective behaviour of larger systems. To quote Richard Dawkins: "If you want to understand life, don't think about vibrant, throbbing gels and oozes, think about information technology."³

Such system-level approaches are emerging in fields as diverse as biology, climate and seismology. A frequent goal is to develop high-fidelity computer simulations as tools for studying system-level behaviour. Computer science, as the 'science of complexity', has much to say about how such simulations — which can be considered a new class of experimental apparatus — should be constructed, and how their output should be analysed and compared with experiment⁴. Similarly, information theory provides formidable insight



The sciences rely on computers, but the benefits are two-way; each is driving the other forwards.

into how biological systems encode, transform and transmit information.

Both the data deluge and system-level science demand computing technology in all its forms — hardware, software, algorithms and theory. The growing importance of computing has several implications for the science of 2020, of which I explore three here.

First, the scientist of 2020 will be adept in computing: not only will they know how to program, but they will have a solid grounding in, for example, the principles and techniques by which information is managed; the possibilities and limitations of numerical simulation; and the concepts and tools by which large software systems are constructed, tested and evolved. This knowledge has been picked up on the job by many pioneering scientists and will hopefully be instilled in the next generation by more formal training. The idea that you can be a competent scientist without such training will soon seem as odd as the notion that you need not have a solid grounding in seventeenth-century mathematics (such as algebra).

Fruitful partnerships

Second, successful science collaborations of 2020 will include computer scientists as key members. All scientists will be adept at applying existing computational techniques, but they will also understand that progress in their fields will require innovation in computing technology. So they will work with computer scientists to identify computational problems, much as today's experimentalists and theorists understand the strengths and weaknesses of their favoured methods and know to partner with others when new techniques are needed. Indeed, this fruitful interplay is

already occurring: for example, the communication challenges inherent in far-flung physics collaborations inspired the development of the World Wide Web, and the need for efficient indexing of terabytes of digital astronomy data has spurred new approaches to organizing spatial data in relational databases⁵. Elsewhere, the scientific opportunities that arise from using wireless sensor networks for, say, continuous habitat monitoring are driving innovations in network protocols and algorithms.

Third, the scientific disciplines and institutions of 2020 will need to train, attract and reward researchers whose focus is on producing the computing innovations required for science to advance: what we might term 'applied computing'. Thus, we see new organizational structures, such as the Computation Institute at the University of Chicago and Argonne National Laboratory (for which I work) and Harvard's Institute for Innovative Computing, that aim to bridge the distinct concerns of computer science and other sciences. Academic departments are hiring faculty with strong computational inclinations. National laboratories have established computational directorates. It will be interesting to see which of these interdisciplinary structures work best.

The growing importance of applied computing also has implications for computer science. Indeed, just as during the early days of the sciences, scientific concerns drove mathematics forward (think of the origins of the calculus), so the many challenging problems posed by modern science can help to focus and motivate research in computing. In my view, it is no accident that some of the most vibrant areas in computing today are those tightly coupled to scientific problems. These dynamics occur in fields as diverse as sensor networks, data integration and grid computing. It's a two-way street, and always has been. ■

Ian Foster is director of the Computation Institute at the University of Chicago and Argonne National Laboratory, Illinois, USA.

Acknowledgements To Anne Rogers for her comments.

1. Denning, P. et al. *Commun. ACM* **32**, 9–23 (1989).
2. Djorgovski, G. S. *Virtual Astronomy, Information Technology, and the New Scientific Methodology*. 7th Int. Workshop Comput. Arch. for Machine Percept. 125–132 (2005).
3. Dawkins, R. *The Blind Watchmaker* (Norton, 1986).
4. Cohen, J. *PLoS Biol.* **12**, 2017–2023 (2005).
5. Szalay, A. S. et al. *SIGMOD Rec.* **29**, 451–462 (2000).

BOOKS & ARTS

When computers take over

What if the current exponential increase in information-processing power could continue unabated?

The Singularity is Near: When Humans Transcend Biology

by Ray Kurzweil

Viking/Duckworth: 2005. 672 pp.
\$29.95/£14.99

Paul Davies

I recall reading a statistic from my student days to the effect that if physics journals continued to grow at the same rate, then by the end of the twentieth century, library bookshelves would have to expand at the speed of light to accommodate them. This absurdity is an illustration of what one might call the exponential-growth fallacy. Examples drawn from technology are legion. The Moon landing in 1969 was widely touted as the first small step on an escalator to the stars, with Arthur C. Clarke predicting huge lunar bases and a Jupiter expedition by 2001. The rapid uptake of robotics in the manufacturing industry after the Second World War led to predictions of cyborg servants and android armies within a few decades. In the event, these technologies became stuck or even slid backwards.

The key point about exponential growth is that it never lasts. The conditions for runaway expansion are always peculiar and temporary (with the possible exception of the expanding Universe). But this sobering fact has not stopped futurologist and author Ray Kurzweil from invoking exponential, and even hyper-exponential, growth in the realm of information processing. There is no doubt that the rise and rise in computing power has dazzled us all. Gordon Moore, co-founder of Intel, famously predicted about 30 years ago that computer processing power would double every 18 months, and so far his prediction has come true. Kurzweil invokes 'Moore's law' as if it were a law of nature, and extrapolates from it into a not-so-distant future in which burgeoning information processing transforms and transcends life as we know it. He refers to his culmination point, at which familiar human culture is obliterated by a tidal wave of unrestrained computation, as the 'singularity'. The word is loosely analogous to the mathematician's singularity, at which the rate of change of a quantity becomes infinite.

If the sky's the limit when it comes to processing information, it isn't hard to think of startling applications. Tired of deciding what to eat? Let a swarm of sensors patrol your



A giant leap: will increased computing power allow machines to take over, as in films like *The Matrix*?

innards and order the right nutrients automatically through your personal wireless network. Concerned about dying? Then achieve immortality by flooding your body with smart nanobots to monitor and maintain your failing biosystems. Or better yet, 'upload' your mind into cyberspace, where you can have more fun, untrammelled by a material body.

The question is not whether these wild ideas should be taken seriously, but whether the premise on which they are founded — unbounded exponential growth in information-processing power — somehow escapes the strictures that eventually curtail all other cases of headlong expansion. One obvious reason why accelerating growth in computation might stall concerns the availability of resources. What happens when the Earth's entire surface has been converted into a gigantic information-garnering, bit-churning system? Kurzweil is ready with the answer: we move into space. ('We' here is a generic term. The sentient beings that will soon wrest control from humans, and which are destined to supervise the cosmic phase of development, will be some sort of superdupercomputers.) But by the remorseless logic of exponentiation, pretty soon thereafter the resource-hungry system will find itself spreading across the galaxy so fast it hits the speed of light. Like the physics journals on

the bookshelves, exponential growth stops here. Or does it? Kurzweil toys with the idea that the speed-of-light barrier is there to be broken, which opens up the giddy prospect of the entire Universe being taken over by an omniscient superintelligence within just a few centuries.

Such exhilarating speculation is great fun to read, but needs to be taken with a huge dose of salt. The biggest lacuna in Kurzweil's argument is the tacit assumption that if we liberate enough information-processing power, then nature will succumb to all our desires. Control the Solar System? Just double the bit rate a few times and it will be within our grasp. Create life? Simulate consciousness? It all boils down to making a cheaper, faster processor. Unfortunately, the laws of physics may well dictate otherwise. Technology can harness physical laws but it can't bend them. No amount of information processing will suspend the law of gravity or create perpetual-motion machines.

When it comes to discussing the physics that underpins his predictions, Kurzweil is apt to be vague or even misinformed. The stupendous power demands implied by the rampant growth in computation and nanotechnology will be met by a concomitant 'law of accelerating returns' in power-generation technologies, such as fuel cells and high-temperature superconductors. And if these run into technical

RONALD GRANT ARCHIVE

problems, well, just send in the nanobots to sort them out. "All technologies," claims Kurzweil, "will essentially become information technologies, including energy."

On the vexed issue of the speed of light, Kurzweil cites evidence that the fine-structure constant, which expresses the strength of the electromagnetic force and contains the speed of light as a factor, may have increased very slightly over the past 6 billion years. The primary evidence comes from an analysis of quasar spectral lines by John Webb of the University of New South Wales in Australia and his collaborators, not from a study of the Oklo natural nuclear reactor in Gabon, as Kurzweil states. Furthermore, even if the observations opened the way to manipulating the value of the fine-structure constant, that is not the same as increasing the speed of light and leaving

everything else unchanged. Indeed, the manipulation would involve a reduction of the fine-structure constant, which would slow the rate of information processing at the atomic level, and so prove self-defeating.

These technical hiccups are irritating, but the book should not be read as a scientific treatise. Rather, it is a futuristic and somewhat breathless romp across the outer reaches of technological possibility, limited only by human imagination. Kurzweil coins the horrible term 'singularitarian' for someone who embraces his vision with alacrity. If Kurzweil is to be believed, we will all be singularitarians in just 29 years' time. Hang in there. ■

Paul Davies is at the Australian Centre for Astrobiology, Macquarie University, Sydney, New South Wales 2109, Australia. His latest book is *How to Build a Time Machine*.

poor were often given, or sold cheaply, animals deemed unfit for commercial consumption.

Social fears often generate fantasies of a golden age, when football hooliganism was unknown, crime was low, nuclear families feasted on wholesome food, and everyone got on well with their neighbours. There is none of that in this volume: the poor struggled, ate badly and died young. Ferrières argues convincingly, I think, that the nineteenth century created the modern food dilemma. Canning, transportation and, eventually, refrigeration extended the food chain. People who would traditionally have seen, touched and smelled the food they were buying now needed only to read the labels. Instead of trusting themselves, they had to trust others. Food inspection declined as political and economic liberalism grew. At the same time, more widespread affluence put a premium on taste, often at the expense of safety. Miasmatic theories of disease and bourgeois sensibilities ensured that the slaughtering was done outside the city centres. Animals no longer had to be able to walk to their destinies.

Ferrières concentrates primarily on France, although there are insights into other cultural experiences, including the persisting British policy of stamping out disease using wholesale slaughter, which is a clumsy tool of disease control. She has a fine sense of the dramatic, faithfully conveyed in the translation, which is sometimes literal at the expense of easy fluency. Away from France, some of the details become fuzzy: a London suburb appears as Issington, rather than Islington, and George Barker has transmuted into someone called Thomas Halwek.

This book is not for the squeamish, but those interested in the culture of the table, or the historical intersections of health, taste and diet, will find plenty here to satisfy their appetites. ■ W. F. Bynum is at the Wellcome Trust Centre for the History of Medicine, University College London, London NW1 2BE, UK.

A taste of a rotten past

Sacred Cow, Mad Cow: A History of Food Fears

by Madeleine Ferrières

Columbia University Press: 2005. 416 pp. \$29.50

W. F. Bynum

Among the formidable list of modern anxieties, those related to food safety loom large. Mad cow disease, genetically modified crops, factory farming, residual pesticides, additives — we read about them daily and are reminded of them when we shop by labels assuring us that we don't need to worry when buying this particular product. We are, after all, what we eat; or more to the point, we are defined by what we do not eat.

All these contemporary concerns give Madeleine Ferrières' monograph a powerful topicality. To her credit, she never panders to the worried well, but sticks to a rich and well-exploited range of historical sources: advice manuals, legal records, statutes, court cases, medical textbooks and imaginative literature. Her running theme is boldly stated: food fears are perennial, although they take different forms depending on the cultural milieu. But the subsidiary theme is also amply demonstrated: fear about food quality varies inversely with fear about food quantity, both temporally and through social strata. If your belly is empty, you don't worry too much about additives or saturated fatty acids.

The title of the original French edition highlights the book's chronological coverage, which ranges from the Middle Ages to the beginning of the twentieth century. So there is nothing about mad cow disease here, despite the allusion in the title of the English version. Instead, in earlier generations, domestic animals with tuberculosis, trichinosis, foot-and-mouth disease and a variety of other ailments made their

weary way to the butchers, and thence to dinner tables. Many would have been especially weary, having been driven hundreds of miles to urban markets. If they couldn't make it under their own steam, they were not permitted to be sold. Those that could walk the last mile would be inspected, slaughtered and consumed quickly. The idea of allowing meat to age after slaughter is a recent luxury. So is the tendency to eat younger animals, rather than those past their capacity to provide milk or wool.

Ferrières also considers other foodstuffs, notably fruit, vegetables and grain. These invoked their own fears, especially unripe fruit and mouldy bread, but most legislation revolved around meat, which was a central part of diets half-a-millennium ago. Economic circumstance determined whether one ate tripe or rotten meat with one's bread, not whether vegetable stew was the main dish. The

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Food to die for? Good food was available to previous generations, but not all of it reached this standard.

Boning up on dinosaurs

The Microstructure of Dinosaur Bone: Deciphering Biology with Fine-Scale Techniques

by Anusuya Chinsamy-Turan

Johns Hopkins University Press: 2005.

216 pp. \$85, £56.50

Luis M. Chiappe

Studies of the microstructure of dinosaur bone began soon after the first scientific reports of these Mesozoic behemoths, but it took nearly 150 years for this discipline to enter the mainstream of dinosaur palaeontology. Nowadays, these studies are used to infer numerous aspects of dinosaur biology, from growth rates and physiological strategies to skeletal biomechanics and developmental patterns, and are even used as a source of new characters for phylogenetic inference. *The Microstructure of Dinosaur Bone* by Anusuya Chinsamy-Turan is the first book-length review of this rapidly advancing line of palaeontological research.

Chinsamy-Turan begins by introducing a host of topics, including the basics of bone composition, histological studies, the changes that occur on death and fossilization, and dinosaur evolution and ecology. Next, she examines the biological meaning of bone microstructure. She looks at bone-tissue dynamics (the rate at which it is deposited, and whether it occurs without interruption or with growth rings) and vascularization (the number of blood vessels), and explains how they can provide clues to the ontogenetic, biomechanical and ecological contexts of bone.

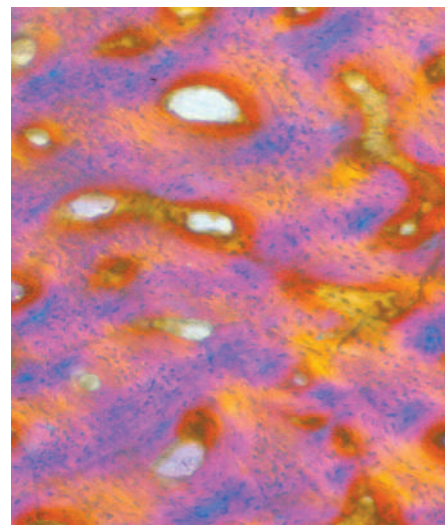
After that she turns her attention to dinosaurs, with a brief history of dinosaur-bone histology and a review of the diverse tissues found in their bones. We then reach the heart of the book, with a discussion of how microstructural studies of fossil bone inform our understanding of the growth patterns of extinct vertebrates. Particularly useful is the author's emphasis on the caveats for inferring some aspects of growth dynamics from fossil bones. Then she covers a number of studies, mostly her own, on the bone histology of pre-modern, Mesozoic birds — dinosaurs' most immediate descendants — and the important changes in growth patterns thought to have occurred during their evolution. Finally, she delves into dinosaur physiology to see if bone microstructure can tell us whether some, all or none of these animals were warm-blooded.

The book draws on Chinsamy-Turan's extensive studies of the bone microstructure of living tetrapods. It is perhaps her deep understanding of the complexity of skeletal growth that leads her continuously to warn against making a strong connection between the different types of tissue preserved in fossilized bones and either specific rates of bone formation or dinosaur physiology. Throughout the

book, Chinsamy-Turan is resolute in defending her views from opposing perspectives; at times it feels like we are watching a television commercial for a video-game. Fortunately, however, the different contenders in the modern arena of dinosaur-bone microstructure agree on several fundamentals. Thanks to work by Chinsamy-Turan and her colleagues and mentors, we know that even if the growth strategies of dinosaurs varied greatly, these animals grew at rates substantially faster than extant non-avian reptiles (even the largest dinosaurs reached adult size in less than three decades). We have also learned that the same rate of bone formation may lead to the creation of different types of bone tissue, and that growth series and data from different skeletal elements are critical for confidently inferring the growth patterns of extinct dinosaurs.

Not surprisingly, the area of greatest disagreement is the largely conjectural issue of whether, in a physiological sense, dinosaurs more closely resembled cold-blooded reptiles or warm-blooded birds and mammals. Most researchers agree that rapid growth cannot be confidently correlated with warm-bloodedness, and most of the disagreement seems to be akin to whether a glass is viewed as half full or half empty. Chinsamy-Turan highlights how the bone microstructure of dinosaurs shows features that are either frequently found in non-avian reptiles or widespread among birds and mammals, but she wisely advises us not to draw definitive physiological conclusions from bone microstructure.

Chinsamy-Turan is at her best when she discusses the microstructure of bone and its



A. CHINSAMY-TURAN

Growth area: microstructure study reveals stages of bone formation in the dinosaur *Dryosaurus*.

significance for understanding the biology of fossil vertebrates. However, her conclusions about the linkage between the origin of warm-bloodedness and the divergence of ornithurines (the group including *Hesperornis*, *Ichthyornis* and all living birds) are weakened by the endorsement of views that have a morphological and functional basis that is not clearly substantiated. Despite these shortcomings, Chinsamy-Turan's book is a tour de force that gives readers an updated synopsis of a fast-growing field of vertebrate palaeontology. The book is a must-read for anybody interested in the biology of one of the most fascinating animals in the history of our planet. ■

Luis M. Chiappe is at the Dinosaur Institute, Natural History Museum of Los Angeles County, 900 Exposition Boulevard, Los Angeles, California 90007, USA.

Summing up physics

The Equations: Icons of Knowledge

by Sander Bais

Harvard University Press: 2005. 96 pp.

\$18.95, £11.95, €17.50

Malcolm Longair

Perhaps the most common comment I receive after delivering a public lecture is: "I enjoyed the lecture, but I couldn't understand the maths." This difficulty doesn't just affect the public but is one that students and scientists encounter throughout their careers: how does one connect the behaviour of the real world with the mathematics used to describe it? After years of practice and hard thinking, professionals become adept at making this connection, which I regard as perhaps the greatest challenge in teaching mathematical physics.

Sander Bais's little book *The Equations* tackles this problem from the point of view of 17 of

the basic sets of physics equations, which the subtitle refers to as "icons of knowledge". In a book of only 96 pages, it is a real challenge to do much more than indicate to the reader the enormous richness of these equations and the imaginative ways they can be used to extend our understanding of the workings of nature at all levels. The equations are presented in their final definitive forms without any discussion of how they came about or why they have these forms. This latter aspect is not part of Bais's agenda, which is a pity. Still, if it were, the book would have grown out of all proportion.

In the first few pages there is a lightning review of elementary mathematical operations, and this helps the reader understand the simpler equations, such as Newton's laws of motion and the continuity equation. It is debatable, however, whether these insights enable the reader to understand the importance of

lagrangians, tensors, spin matrices and so on. As the equations progress through the Dirac equation, quantum chromodynamics and electroweak theory to the superstring action, Bais simply writes down the equations in their most compact and elegant form and discusses what the solutions mean. This is fair enough if we are to regard the equations simply as icons of knowledge.

But when reading the book, I kept thinking: "If only he had said..." For example, why are complex numbers the natural language of quantum mechanics? Where do the horrendous complexities of turbulence come from in the Navier-Stokes equations? What was the step of genius that led James Clerk Maxwell to the final form of his equations for the electromagnetic field? If the key concept of the relativity of simultaneity had been introduced, so many of the problems of special relativity

would have disappeared. For me, these would have added to the insight and value of the exposition. However, this is not really what the book is about — it is much more a concise exposition of what the equations can do, and the reader has to trust that the author has got it right. Indeed I cannot fault what he says about the meanings of the equations, and the main text is enlivened by brief sketches of the lives of some of the principal players. I only wish he had included Galileo's pivotal contributions as the founder of the whole business: "The book of Nature is written in mathematical characters."

Who will gain most from reading this book? It has to be someone who wants an introduction to the power of mathematics in describing natural phenomena without actually having to do the maths. The average interested public should be encouraged to dip into the book and

see how they get on. My suspicion is that the most important target audiences are young people who aspire to become mathematically oriented scientists and who will find the scope of the book inspiring. It would be a lovely birthday gift for them.

At a recent public lecture on Einstein and the arts in the first decades of the twentieth century, my co-presenter from the perspective of art history commented on the sheer physical beauty of the equations that I risked presenting to the audience. In addition to its pedagogical value, Bais's book presents these icons of our physical world in all their beauty. It is very good to be reminded of this. ■

Malcolm Longair is Jacksonian professor of natural philosophy, Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK. His book *The Cosmic Century* will be published in April 2006.

Porcelain perception

Not everything is as it seems in the ceramics of Pauline Wiertz.

Colin Martin

Contemporary artists sometimes suppress any national or historic references in their work, perhaps to make it more universal and hence more appealing on the global fine-art market. Whatever the reason, much of the work exhibited at Collect, an international contemporary art fair held in London last month, was aesthetically unadventurous and almost 'stateless' in its lack of cultural interest.

One outstanding exception was the covetable slip-cast porcelain and earthenware exhibited by Dutch ceramist Pauline Wiertz. Unashamedly relishing her national artistic heritage, her recent work evokes the sumptuous still-life canvases of heaped fish, meat, fruit and vegetables familiar from the seventeenth-century 'golden age' of Dutch painting.

At first glance, *Garnalen Cocktail* ('shrimp cocktail') seems to be a typical assortment of marine species. Life-like pink prawns are heaped on lustrous black-glazed shells and starfish, and the piece is given height by an armature of what appears to be black coral prettily tipped with rose pink at the rear. On closer inspection, however, it becomes clear



Neither fish nor fowl: Pauline Wiertz uses chicken feet in *Garnalen Cocktail*.

that Wiertz used severed chicken feet to cast the coral, although the shellfish were cast from real specimens.

This playful hoax echoes the contents of a seventeenth-century *Wunderkammer*, or chamber of wonders, a cabinet filled with curiosities by a wealthy, private collector. Although a *Wunderkammer* typically housed preserved animals, skeletons, horns and tusks, it might also include man-made

oddities that mixed fact with fiction. The cabinet of Ole Worm (1588-1654), who taught medicine at the University of Copenhagen in Denmark, contained a woolly fern masquerading as a 'Scythian lamb', thought to be a hybrid of a plant and a sheep.

In its verisimilitude, Wiertz's work resembles the eighteenth-century rococo porcelain produced by the Meissen factory in Germany. By using chicken feet to cast moulds for her ceramic coral, Wiertz encourages us to think about the nature of perception: if we expect to see coral in the context of other marine species, then that is what we will discern. There is also the hint of a wry comment on our contemporary preoccupation with the

spread of the H5N1 strain of avian flu. This point is made more clearly in her appositely named *Kippenpootjes* ('chicken legs' or 'chicken fever'), a series of individual chicken feet slip-cast in porcelain and colourfully glazed.

Wiertz's ceramics can be seen at the Galerie Terra Delft in the Netherlands from 25 March to 22 April.

Colin Martin is a writer based in London.

R. ZILUSTR

MOLECULAR BIOLOGY

RNA lost in translation

David Tollervey

In any manufacturing process, quality control is crucial, and gene expression is no exception. A new pathway monitors mRNAs — the intermediaries of gene expression — and destroys faulty molecules.

For most human genes to be usefully expressed they must first be copied into messenger RNAs by the process of transcription. These then program large, RNA–protein complexes called ribosomes to synthesize a specific protein by ‘translation’. The fidelity of mRNA translation into protein is vital for the overall accuracy of gene expression, and cells have evolved ways to detect any aberrant mRNAs that have structural defects. Doma and Parker (page 561 of this issue)¹ have discovered an mRNA quality-control system — ‘no-go decay’ — with some unexpected features.

Cells alter their rates of mRNA transcription to change mRNA levels, and so rates of protein synthesis, in response to many stimuli. To adjust mRNA levels, cells must be able to rapidly get rid of normal mRNAs that were previously synthesized (turnover). In fact, different mRNAs differ radically in their rates of degradation, and this is subject to both metabolic and developmental regulation. In addition, cells must guard against the synthesis of abnormal mRNAs (surveillance), which can produce defective, potentially toxic, protein products.

The normal turnover of mRNAs has been extensively analysed in budding yeast. In this organism, two main pathways have been identified, each of which degrades the mRNA from only one of its ends — either the 5′ or the 3′ end. Two further pathways deal only with quality control, specifically degrading mRNAs with structural defects. ‘Nonsense-mediated decay’ recognizes mRNAs where the site of translation termination occurs too early in the RNA. By contrast, mRNAs that entirely lack a translation stop site are subject to ‘non-stop decay’ (reviewed in refs 2, 3).

In each of these four pathways — at least in yeast — the mRNAs are degraded exclusively from their ends. But the no-go decay pathway identified by Doma and Parker¹ is different: this pathway begins its work in the middle of the mRNA (Fig. 1).

The authors engineered an obstruction into mRNAs (a very stable ‘stem-loop’ structure) to stall the translating ribosome, and showed that this triggered degradation. The no-go decay is

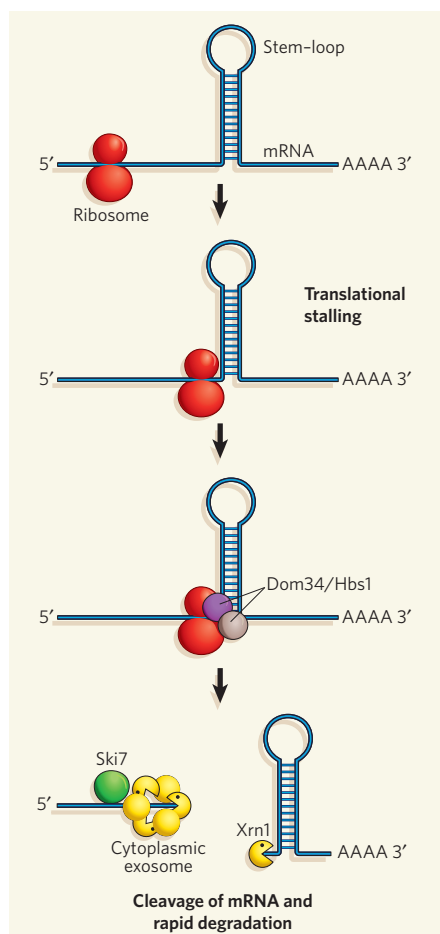


Figure 1 | No-go decay. Doma and Parker¹ show that the no-go decay pathway results in rapid mRNA cleavage and degradation. This pathway is different from other mRNA degradation pathways in yeast, in that stalling of the ribosome by a stem-loop structure leads to cleavage of the mRNA¹, a process that involves the Dom34 and Hbs1 proteins⁴. This cleavage generates free ends that are subject to degradation: the enzyme Xrn1 chews up the mRNA from the 5′ end, whereas the multi-enzyme exosome complex degrades from the 3′ end. Dom34 and Hbs1 might interact directly with the ‘A site’ of the stalled ribosome, possibly to promote release of the ribosome and allow subsequent mRNA cleavage, or they may directly stimulate cleavage of the mRNA.

initiated by cleavage of the mRNA close to the stall site. This break generates entry sites for enzymes that degrade the remainder of the RNA: the free 5′ end is degraded by Xrn1, and the 3′ end is degraded by a complex termed the exosome (Fig. 1). Both enzymes have similar roles in the other degradation pathways. No-go degradation definitely requires stalling of the ribosome, because the termination of translation before the obstruction prevents cleavage of the mRNA¹.

The initial cleavage of the mRNA involves two proteins called Dom34 and Hbs1. These proteins are related to translation factors⁴, which regulate translation through direct contact with the ribosome. So it looks as though Dom34 and Hbs1 might interact with the stalled ribosome. Hbs1 is also related to a protein called Ski7 that is an accessory to the exosome⁵. During non-stop decay, Ski7 is proposed to release ribosomes stalled at the ends of mRNAs that lack a translation-termination site⁶. Ski7 is also needed when the exosome degrades the cleaved no-go mRNAs¹.

How the mRNA cleavage occurs in no-go decay has yet to be worked out. Obvious possibilities are that Hbs1 or Dom34 catalyse it themselves, although their sequences do not suggest this, or they could recruit a specific nuclease enzyme to the site. In either case, it would presumably be necessary to first remove the stalled ribosome, because inspection of the end points of the cleaved RNA does not show an obvious ‘footprint’ — an area protected from cleavage by the bound ribosome.

Alternatively, interaction with Hbs1 or Dom34 might allow the ribosome itself to cleave the mRNA. In support of this, in the bacterium *Escherichia coli* pausing of translation can cause mRNA cleavage by the ribosome (at the ‘A site’ of the ribosome)^{7,8}. The cleavage converts the mRNA into a non-stop transcript. This can then be dealt with by the bacterium’s non-stop decay pathway, which releases the stalled ribosome and degrades the problematic mRNA.

It is attractive to think that a similar mechanism might operate in yeast. However, the ends of the fragments that Doma and Parker¹

detected following no-go cleavage were quite heterogeneous — more so than might be predicted by cleavage within the stalled ribosome. But in other contexts, mRNA fragments stabilized by the absence of Xrn1 or the exosome are also heterogeneous^{9,10}. So it could be that the free ends are nibbled away by other nucleases after the mRNAs have been cleaved by the ribosome.

It is not clear how widespread no-go decay is in yeast, or whether it is conserved in other animals. In fruitflies, recognition of mRNAs by the nonsense-mediated decay system also leads to their cleavage in the vicinity of the translation stop site¹¹. This triggers rapid degradation by Xrn1 and the exosome, hinting that a mechanism similar to no-go decay might operate in the flies.

Notably, Doma and Parker report substantial cleavage of mRNAs containing an artificial stem-loop structure that should create an almost impassable block for the ribosome. But other, perhaps more physiologically relevant, mRNAs with translational pause sites were much less subject to no-go decay. It may be that the main targets for no-go decay in the cell are chemically damaged mRNAs, which can cause a complete translation block¹².

In all organisms, the amount of each mRNA is very tightly controlled — a crucial step in the regulation of gene expression. In Doma and Parker's experiments, the no-go pathway is apparently acting specifically as a quality-control system. However, it seems likely that in other contexts this pathway will have been co-opted to regulate the abundance of specific mRNAs. Consistent with this possibility, in mice and fruitflies a mutation of a close relative of Dom34 called Pelota leads to specific defects in development and cell-cycle progression^{13,14}.

The integrity and functionality of mRNAs are monitored by surveillance activities that act at many steps during the molecules' nuclear maturation and cytoplasmic lifetime. Such continuous surveillance to maintain high fidelity is metabolically expensive but, as elsewhere in life, infidelity is surely still more costly. ■

David Tollervey is at the Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3JR, UK.
e-mail: d.tollervey@ed.ac.uk

BIOGEOCHEMISTRY

Gas with an ancient history

Don E. Canfield

Researchers persist in tackling our ignorance of what life was like way back in Earth's history. Evidence of methane production in ancient microbial ecosystems now emerges from 3.5-billion-year-old rocks.

Piecing together the early history of life from the geological record is like putting together an incomplete jigsaw puzzle. A few of the pieces connect, but many of them don't, and most are missing. We know — or at least we think we know — that life evolved by 3.8 billion years ago. The evidence, found in the isotopic composition of organic carbon¹, is rather scarce, but it is compelling. Carbon isotopes of this type, however, cannot tell us which organisms were present.

For further insights into early Earth ecosystems, we zoom forward around 300 million years to the Dresser Formation in Western Australia. It is from here that Ueno *et al.* (page 516 of this issue)² report evidence for the existence of an ancient population of methane-producing microbes, commonly known as methanogens. This represents one of the earliest specific microbial processes to have been identified in the geological record: it adds an extra piece to the emerging picture of the structure and complexity of ancient microbial ecosystems.

The story is this. Between 3.5 billion and 3.45 billion years ago, a series of lavas known as pillow basalts, and sedimentary rocks, were deposited in what is now the Dresser Formation. The basalts were intruded by numerous silica dykes, some of which terminate in the sedimentary layers (see Fig. 1 of the paper on page 517). This means that the dykes are ancient and were formed during the deposition of the sediments. Small bubbles (fluid inclusions) are found within quartz minerals in the dykes, and these bubbles contain mainly carbon dioxide and water, but also some methane (CH₄). Many of the bubbles align along the growth margins of the quartz, and so are believed to have formed as the quartz was precipitated: they are viewed as primary inclusions. Other bubbles are scattered randomly within the quartz (and are probably primary), while still others are concentrated where the quartz was fractured and resealed. These are viewed as secondary inclusions.

In their investigations, Ueno *et al.*² crushed samples of quartz (Fig. 1) and measured the concentrations of the extracted gases and the isotopic composition of the methane carbon. The primary and secondary inclusions contained methane with distinct isotopic compositions. The secondary inclusions had a composition consistent with (but not proof of) a non-biological, thermal source of methane that was probably introduced some time after

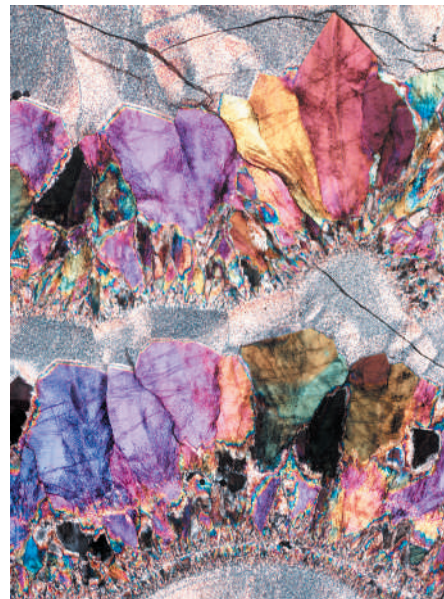


Figure 1 | Inclusive evidence. This optical photomicrograph shows quartz from the Dresser Formation in cross-polarized light. The coloured crystals are coarse-grained quartz, which contain the tiny fluid inclusions analysed by Ueno *et al.*². (Image courtesy of Yuichiro Ueno; width is about 3 mm.)

the silica dykes were first in place. By contrast, the methane from the primary inclusions was much more strongly depleted in ¹³C, and is consistent with a microbial origin.

Still, the isotopic composition of the methane from the primary inclusions is not very different from that of the most ¹³C-depleted, thermally formed methane. Might all of the methane have a non-biological source? The authors anticipate this argument and note that their primary fluid inclusions lack the higher (C₂₊) hydrocarbons that are normally associated with the most strongly ¹³C-depleted thermogenic methane. All in all, I take the data at face value, and accept that the authors have probably uncovered the oldest-known samples of biologically produced gas.

With this backdrop, let us enquire more broadly into the nature of life 3.5 billion years ago. Evidence for the activities of sulphate-reducing bacteria, which use sulphate rather than oxygen in energy production, has been reported from the Dresser Formation³. Fossil structures known as stromatolites, which are probably of biological origin, have been found in slightly younger rocks from the same stratigraphic sequence⁴. The organisms

1. Doma, M. K. & Parker, R. *Nature* **440**, 561–564 (2006).
2. Conti, E. & Izaurralde, E. *Curr. Opin. Cell Biol.* **17**, 316–325 (2005).
3. Wagner, E. & Lykke-Andersen, J. *J. Cell Sci.* **115**, 3033–3038 (2002).
4. Carr-Schmid, A., Pfund, C., Craig, E. A. & Kinzy, T. G. *Mol. Cell. Biol.* **22**, 2564–2574 (2002).
5. van Hoof, A. *Genetics* **171**, 1455–1461 (2005).
6. van Hoof, A., Frischmeyer, P. A., Dietz, H. C. & Parker, R. *Science* **295**, 2262–2264 (2002).
7. Hayes, C. S. & Sauer, R. T. *Mol. Cell* **12**, 903–911 (2003).
8. Sunohara, T., Jojima, K., Yamamoto, Y., Inada, T. & Aiba, H. *RNA* **10**, 378–386 (2004).
9. He, F. & Jacobson, A. *Mol. Cell. Biol.* **21**, 1515–1530 (2001).
10. He, W. & Parker, R. *Genetics* **158**, 1445–1455 (2001).
11. Gatfield, D. & Izaurralde, E. *Nature* **429**, 575–578 (2004).
12. Ougland, R. *et al. Mol. Cell* **16**, 107–116 (2004).
13. Xi, R., Doan, C., Liu, D. & Xie, T. *Development* **132**, 5365–5374 (2005).
14. Adham, I. M. *et al. Mol. Cell. Biol.* **23**, 1470–1476 (2003).

involved in stromatolite growth are unknown. However, microbial mats have been identified in 3.42-billion-year-old rocks from South Africa⁵, and the chemistry of the rocks indicates that anoxygenic phototrophs (photosynthetic organisms that do not produce oxygen) were probably the dominant mat-forming organisms.

What about oxygen? An aerobic biosphere was not possible before the evolution of oxygen-producing cyanobacteria, and oxygen-producing photosynthesis accomplishes the vast majority of the primary production today. So whether or not cyanobacteria were present is crucial to our understanding of the nature of the biosphere 3.5 billion years ago.

To explore this issue, we return to Western Australia and to the Apex Chert, which is slightly younger than the Dresser Formation. Here, a series of organic remains has been interpreted as possibly being those of cyanobacteria⁶. Unfortunately, a cyanobacterial affiliation cannot be decided by morphology alone, and the interpretation of these remains as biogenic is controversial⁷. We cannot, therefore, be confident that oxygen-producing organisms existed 3.5 billion years ago. But we can be certain that an active biosphere was in place. It was probably quite complex, including many of the components of modern anaerobic ecosystems. Of these, we can identify anoxygenic phototrophs, sulphate reducers and methanogens.

Another intriguing aspect to Ueno and colleagues' research² is its geological setting. The silica dykes housing the methane originated from deep in Earth's crust — maybe as much as 1 kilometre below the surface⁸. This is not the normal bubbly lake or coastal marine sediment usually associated with methanogenesis. These results, therefore, point to an active subsurface biosphere. A deep subsurface biosphere exists today, and by some estimates⁹ it is as massive as the one on the Earth's surface. The results of Ueno *et al.* do not tell us anything about the magnitude of the ancient subsurface biosphere. But the presence of life within Earth's crust, combined with good evidence for life in surface environments, suggest that by 3.5 billion years ago the Earth was teeming with microorganisms.

Don E. Canfield is at the Nordic Center for Earth Evolution and the Institute of Biology, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark.
e-mail: dec@biology.sdu.dk

1. Rosing, M. T. *Science* **283**, 674–676 (1999).
2. Ueno, Y., Yamada, K., Yoshida, N., Maruyama, S. & Isozaki, Y. *Nature* **440**, 516–519 (2006).
3. Shen, Y., Buick, R. & Canfield, D. E. *Nature* **410**, 77–81 (2001).
4. Hofmann, H. J., Grey, K., Hickman, A. H. & Thorpe, R. I. *Geol. Soc. Am. Bull.* **111**, 1256–1262 (1999).
5. Tice, M. M. & Lowe, D. R. *Nature* **431**, 549–552 (2004).
6. Schopf, J. W. *Science* **260**, 640–646 (1993).
7. Brasier, M. D. *et al.* *Nature* **416**, 76–81 (2002).
8. Ueno, Y., Yoshioka, H., Maruyama, S. & Isozaki, Y. *Geochim. Cosmochim. Acta* **68**, 573–589 (2004).
9. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).

MEMBRANE BIOLOGY

Permutations of permeability

William N. Zagotta

The first glimpse into the molecular basis of how sodium ions are transported across cell membranes by ion channels shows that cation-selective channels are variations on potassium channels.

Ion channels are nature's gatekeepers: by allowing certain ions to pass across the membrane and excluding others, these proteins control the electrical state of the cell. Channels selective for sodium ions allow Na⁺ to enter the cell, causing depolarization of the membrane. In the case of neurons, this can trigger the cell to fire an action potential or release a neurotransmitter. Channels selective for potassium ions allow K⁺ to exit the cell, causing hyperpolarization of the membrane and decreasing the cell's electrical excitability¹. The structural basis for K⁺ selectivity is well established, mainly from work on the bacterial K⁺ channel KcsA (refs 2–4), but the basis for Na⁺ permeability has been unclear. In this issue, Jiang and colleagues (page 570)⁵ reveal the first atomic structure of a channel with both Na⁺ and K⁺ permeability, the NaK channel from *Bacillus cereus*.

Given their opposite roles, it is surprising that many Na⁺ channels and K⁺ channels are structurally related. Both are tetrameric (with four subunits) or pseudotetrameric channels with a 'selectivity filter' in the narrowest part of the pore, formed from the P-loop between the inner and outer transmembrane helices (Fig. 1). This motif is shared by many other ion channels that range in selectivity from highly K⁺-selective to highly Na⁺- or Ca²⁺-selective (Fig. 2, overleaf). Highly K⁺-selective channels almost all possess the amino-acid signature sequence TVGYG in their selectivity filter. Together with the hydroxyl group of the threonine (T), the backbone carbonyls from the signature sequence form four K⁺-binding sites in the selectivity filter (Fig. 2). The approximately 1,000:1 selectivity for K⁺ over Na⁺ of these channels arises from the K⁺ selectivity of these sites and mutual repulsion between the K⁺ ions bound to the sites⁶.

The similarity of selectivity filter sequences to the signature sequence, however, breaks down as channels become more Na⁺ and Ca²⁺ permeable (Fig. 2). The hyperpolarization-activated HCN channels exhibit only about a 4:1 selectivity for K⁺ over Na⁺, and the transient receptor potential (TRP) channels and cyclic nucleotide-gated (CNG) channels are equally selective for K⁺ and Na⁺ (refs 7, 8).

Also, as the K⁺ selectivity decreases, a new feature emerges in these channels — Ca²⁺ permeability. Both the TRP and CNG channels are highly permeable to Ca²⁺. Paradoxically, Ca²⁺ also blocks some of these channels at low concentrations. Ca²⁺ permeability is thought

to arise at high concentrations of Ca²⁺ from the binding of multiple Ca²⁺ ions that exhibit mutual repulsion¹. This property is finely honed in the voltage-gated Ca²⁺ channels to produce extremely high Ca²⁺-selectivity.

Jiang and colleagues⁵ reveal that the properties of Na⁺ permeability and Ca²⁺ block also seem to be shared by the NaK channel from *B. cereus*. NaK contains a selectivity filter sequence (TVGD) similar to that of most CNG channels (TIGE). The structure of the NaK channel, therefore, gives us our first glimpse into how differences in the sequence of the selectivity filter can produce varying Na⁺ and Ca²⁺ permeability in these channels.

Overall, the backbone structure of NaK is very similar to that of KcsA. It is composed of four subunits surrounding a central pore. Each subunit has two membrane-spanning helices, M1 (outer helix) and M2 (inner helix), with an intervening P-loop (Fig. 1). The pore is lined by the M2 on the intracellular side, and by the P-loops on the extracellular side.

The NaK structure diverges from KcsA in the selectivity filter. Molecular dynamics simulations implicated the first two binding sites in the selectivity filter of KcsA as the most K⁺-selective sites⁷. In NaK, these two binding sites are replaced by a 'pore vestibule' that can accommodate just a single ion (Fig. 2). This

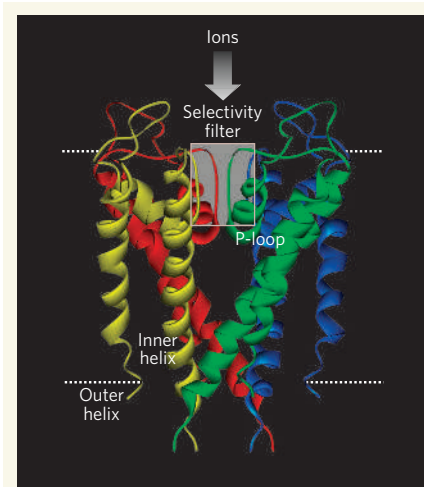


Figure 1 | Structure of the KcsA potassium ion channel. The four subunits of the channel are shown in different colours. They surround a central pore, guarded by the selectivity filter made up of the P-loops from each of the subunits. This filter determines the permeability of the pore to different ions. Dotted lines show the approximate position of the cell membrane.

CHEMISTRY

Perkin, the mauve maker

One hundred and fifty years ago this week, a teenager experimenting in his makeshift home laboratory made a discovery that in effect launched the modern chemicals industry. William Perkin was an 18-year-old student of August Wilhelm Hofmann at the Royal College of Chemistry in London, working on the chemical synthesis of natural products. In a classic case of serendipity, Perkin chanced on his famous 'aniline mauve' dye while attempting to synthesize something else entirely: quinine, then the only known remedy for malaria.

As a student of the renowned

German chemist Justus von Liebig, Hofmann had made a name for himself by showing that a basic compound called aniline, obtained from coal tar, had the same properties as a compound distilled from raw, plant-derived indigo. Coal tar was the residue of coal-gas production, and there was intense interest in finding uses for the aromatic compounds, such as aniline, that could be extracted from it.

Working at home, Perkin tried to use an aniline derivative to make quinine by oxidation, based on the similarity of their chemical formulae (their molecular structures are very

different). The reaction produced only a reddish sludge. But when the inquisitive Perkin tried the reaction using aniline instead, he got a black precipitate that dissolved in methylated spirits to give a purple solution. Textiles and dyeing being big business at the time, Perkin was astute enough to test the coloured compound on silk, which it dyed richly. Hitherto, almost all dyes were natural compounds extracted from plants and animals.

Boldly, Perkin persuaded his father and brother to set up a small factory with him to manufacture the dye, which he called mauve; the picture here is of a shawl, dyed with mauve, from 1862. The Perkins and others (including Hofmann) soon discovered a whole rainbow of



aniline dyes, and by the mid-1860s aniline-dye companies included the nascent giants of today's chemicals industry, such as Bayer, Hoechst and BASF.

Philip Ball

SCIENCE MUSEUM

ion is not bound by carbonyl oxygens, as it would be in KcsA, because of a rearrangement in the backbone of the selectivity filter, presumably caused by the change of tyrosine (Y) in KcsA to aspartate (D) in NaK. The third and fourth ion-binding sites are structurally very similar to their counterparts in KcsA, but can bind to K^+ and Na^+ with virtually no rearrangements in the protein. These differences might make the NaK pore seem shorter in electrophysiological and flux studies than the K^+ channel pore¹.

Another intriguing finding in the NaK structure was the presence of a divalent binding site,

capable of binding Ca^{2+} , at the extracellular entrance to the NaK selectivity filter. Although an extracellular Ca^{2+} -binding site has been proposed in many Ca^{2+} -permeable channels, this site has always been thought to involve the carboxylate of an acidic residue in the selectivity filter. Surprisingly, in the NaK structure this acidic residue, the aspartate, is facing away from the pore, and the binding site for Ca^{2+} is formed by backbone carbonyls of the glycine residues at the extracellular end of the selectivity filter sequence. In CNG channels, this acidic residue has also been proposed to be involved in regulating a number of permeation

and blocking properties, including ion selectivity, divalent block, proton block and state-dependent tetracaine block¹⁰. Although the sequence similarity is weaker between CNG channels and voltage-gated Ca^{2+} and Na^+ channels (Fig. 2), a similar role for an acidic residue has been suggested in those channels¹.

The finding that the aspartic-acid residue in the NaK structure does not bind directly to the ions raises some interesting questions. Can the charge on this residue exhibit a through-space electrostatic interaction with ions in the pore? Do mutations of this residue alter the backbone structure and so indirectly alter ion-binding sites? Could the position of this acidic residue be different in open channels? And, is the NaK channel pore structure representative of eukaryotic CNG channels and voltage-gated Ca^{2+} channels? As always, X-ray crystallography has answered many questions with astonishing clarity, but there are yet more to be addressed.

William N. Zagotta is in the Department of Physiology and Biophysics, University of Washington, Box 357290, Seattle, Washington 98195-7290, USA.

e-mail: zagotta@u.washington.edu

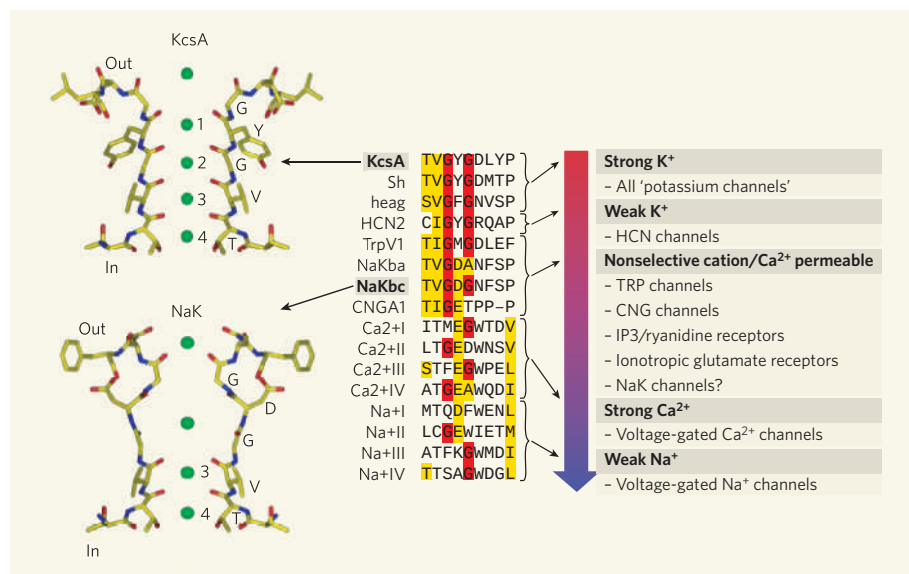


Figure 2 | Mechanism of ion selectivity. The amino-acid sequences from a range of ion channels (centre) show remarkable similarity in the region of the selectivity filter. Red highlights show the conserved glycines in the 'signature' sequence TVGYG found in the potassium ion channel KcsA. Yellow highlights show amino acids that are chemically similar. As the filter sequence becomes less conserved, the channels lose their selectivity for K^+ over Na^+ and become more permeable to Ca^{2+} . The structure of the nonselective cation channel NaK from *Bacillus cereus* has been solved by Jiang and colleagues⁵. Comparison of the selectivity filters of NaK and KcsA (left) provides structural clues to sodium-ion permeability.

- Hille, B. *Ion Channels of Excitable Membranes* (Sinauer, Sunderland, MA, 2001).
- Doyle, D. A. et al. *Science* **280**, 69–77 (1998).
- Morais-Cabral, J. H., Zhou, Y. & MacKinnon, R. *Nature* **414**, 37–42 (2001).
- Zhou, Y., Morais-Cabral, J. H., Kaufman, A. & MacKinnon, R. *Nature* **414**, 43–48 (2001).
- Shi, N., Ye, S., Alam, A., Chen, L. & Jiang, Y. *Nature* **440**, 570–574 (2006).
- Hille, B., Armstrong, C. M. & MacKinnon, R. *Nature Med.* **5**, 1105–1109 (1999).
- Craven, K. B. & Zagotta, W. N. *Annu. Rev. Physiol.* **68**, 375–401 (2006).
- Ramsey, I. S., Delling, M. & Clapham, D. E. *Annu. Rev. Physiol.* **68**, 619–647 (2006).
- Noskov, S. Y., Berneche, S. & Roux, B. *Nature* **431**, 830–834 (2004).
- Kaup, U. B. & Seifert, R. *Physiol. Rev.* **82**, 769–824 (2002).

EARTHQUAKES

A movement in four parts?

Joanne Bourgeois

From time to time, over millennia, northwestern North America has experienced huge earthquakes. These events may be preceded by tell-tale subsidence, but the evidence is devilishly difficult to decipher.

Over the past year, two groups of micropalaeontologists^{1,2} have presented evidence that coastal land subsides not only during a great earthquake but also in the decade or so before such events. Their study areas are the Alaska and Cascadia subduction zones in northwestern North America. A subduction zone is where one tectonic plate dives beneath another, and in Alaska and Cascadia tectonic shifts have given rise to earthquakes of moment magnitude 9 or more (for comparison, the Sumatra–Andaman earthquake that led to the tsunami of 26 December 2004 was moment magnitude 9.2). The observations and interpretations of pre-earthquake subsidence are complex and subtle. But if this subsidence does indeed occur, the implication is that warning signs are detectable for some time before one of these huge earthquakes occurs.

Evidence of subsidence during great earthquakes has been well-documented, but the record between these events is more difficult to tease out. The most complete set of data comes from the shores of Cook Inlet, near Anchorage, Alaska. There and elsewhere along a fault rupture 800 km in length, a moment magnitude-9.2 earthquake in 1964 was accompanied by 1–2 m of co-seismic subsidence. Another source of data is the 1700 Cascadia earthquake, which was probably in the range 8.7–9.2 and produced similar effects, as recorded in coastal marsh stratigraphy. In Alaska, at least five previous earthquakes have been studied, going back about 3,000 years; for Cascadia, eight or more events in the past 5,000 years have been identified.

For these and other subduction-zone coastlines, a three-part seismic–interseismic deformation cycle has been considered standard (Fig. 1). First, between earthquakes the shallow part of the subduction zone is locked, strain accumulates on the fault and the coastline rises (with subsidence offshore). Second, during the earthquake (the co-seismic phase), the uplifted region quickly subsides, while the offshore region undergoes rapid uplift; this rapid motion generates a tsunami. Third, during the interseismic period the earthquake deformation is recovered at first rapidly (the post-seismic phase), and then more slowly (the interseismic phase). This cycle, typically repeating over some hundreds of years, may be superimposed on longer-term patterns of net uplift or subsidence. Local changes in sea level will also occur, on a timescale of years or decades, caused for example by storm cycles, by glacial ‘loading’, by global fluctuations in

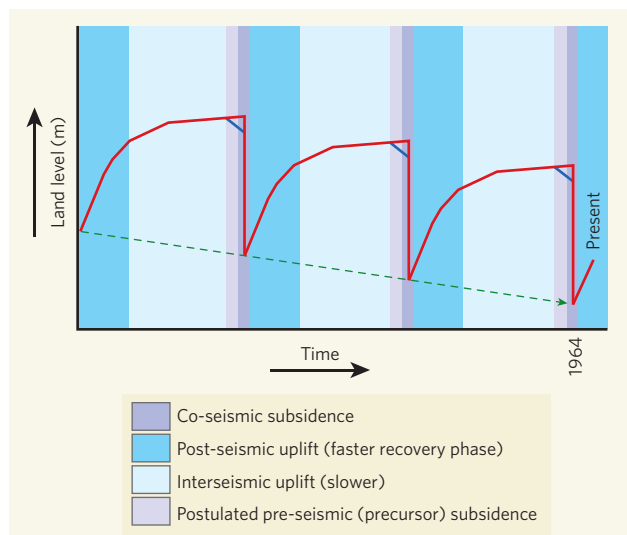


Figure 1 | An idealized seismic cycle for a coastal area that subsides during great earthquakes. The red line shows the accepted cycle of the rise and fall of land; the blue line shows the small-scale subsidence that is postulated to precede great earthquakes; and the dashed green line shows net subsidence over time. The period between the earthquakes is on the order of centuries, with the end of the series showing the present elevation of the Alaska coast following the 1964 subduction-zone earthquake.

sea level, and by variability in local patterns of crustal stress.

In the mid-1980s and the 1990s, initial studies of the subduction-zone deformation cycle in coastal sedimentary records focused on dramatic changes in environment associated with co-seismic subsidence — for example, the sharp elevation drop evident in a change from marsh peat laid down above high tide to mud deposits from the intertidal zone. Beginning in the mid-1990s, Ian Shennan and colleagues^{3–5}, among others, began to look at microfossils in these sedimentary records in Alaska and Cascadia (different microfossils reveal different habitat, be it sediment soaked in salt, brackish or fresh water). The aim was to quantify the amount of co-seismic subsidence (typically 1–2 m) that occurs, and also to distinguish earthquake-generated change from other causes as seen, for instance, along aseismic coastlines. In doing so, Shennan and his co-workers also found subtle evidence of other effects. In some cases, as seen at Johns River in Washington state, the expected uplift continued until the co-seismic subsidence. In other cases, however, there was a small amount of pre-seismic subsidence (Girdwood Flats, Alaska). And in other cases again, the evidence was equivocal (Netarts Bay, Oregon).

Shennan's team identified key species of microfossils for diagnosing the magnitude of subsidence, and developed methods for quantifying the amount of elevation change using a mathematical approach involving transfer functions. As work continued, they found several additional cases of apparent pre-seismic subsidence of the order of 10–30 cm ± 10–20 cm.

Only for the 1964 case in Alaska has the duration of this signal been estimated — about a decade, determined from concentrations of caesium-137, an isotope produced by bomb testing⁶.

Based on some of these analyses, Shennan *et al.*⁵ proposed a four-part deformation cycle, adding this short period of decimetre-scale subsidence before the main event (Fig. 1). The new work by Shennan and Hamilton¹, as well as analyses by Hawkes *et al.*² that include some new microfossil types, support the four-part cycle for many but not all cases of great earthquakes in Alaska and Cascadia.

The Alaskan microfossil records also imply the occurrence of several episodes of slight uplift and subsidence that lack obvious association with earthquakes. Shennan and Hamilton¹ attribute these ‘non-precursor’ episodes to changing sea levels caused by glacier growth and retreat. In contrast, for the precursor events, the authors argue that their position directly below co-seismic events, at several localities, and over some millennia of record, rules out other transient causes such as increased storminess and subsidence beneath the load of a glacier. The authors cite several other kinds of observation that both support and conflict with their hypothesis of precursor subsidence.

The records are subtle — can centimetres to decimetres of vertical change be reliably detected? Can mixing of the microfossil record be ruled out, as the authors claim? Have all alternative explanations been eliminated, for example the possibility that rapid sedimentation following co-seismic subsidence better

preserves the last few storm deposits before the earthquake? Are there enough analyses to rule out coincidence; that is, do all areas undergo both transient uplift and subsidence on this scale, during the interseismic period, and some are just caught at the time when a significant marker (a large earthquake) occurs?

But if pre-seismic subsidence does occur, what might be the mechanisms? A hypothesis suggested by both groups^{1,2} is the occurrence of 'slow earthquakes' along a deeper part of the subduction zone than the part that ruptures during great earthquakes. Such slip⁷ recurs about every 14 months at Cascadia⁸. But theory has not predicted and observation has not documented subsidence at coastal locations from these slow earthquakes. Moreover, in Cascadia these 'earthquakes' have occurred since the Global Positioning System (GPS) network was installed in 1992, and they appear to happen regularly⁸. Indeed, if the authors' hypothesis is correct, it would predict an imminent great earthquake in Cascadia.

A good test of the idea might have been possible if the Sumatra region had been intensively monitored in the decade before the massive earthquake that occurred on 26 December 2004 — but hindsight is a wonderful commodity. Some parts of coastal areas

along subduction zones have been instrumented with networks of continuous GPS sites in the past 10–15 years. But few areas in the world yet have precise and dense enough coverage with geodetic arrays to detect slow earthquakes; measuring vertical motions is even more of a challenge. Nonetheless, those systems that are in place are helping us to understand the fundamental mechanics of subduction zones. Whether this behaviour predictably includes precursor coastal subsidence remains an open question. ■

Joanne Bourgeois is in the Department of Earth and Space Sciences, University of Washington, Seattle, Washington 98195-1310, USA.
e-mail: jbourgeo@u.washington.edu

- Shennan, I. & Hamilton, S. *Quat. Sci. Rev.* **25**, 1–8 (2006).
- Hawkes, A. D., Scott, D. B., Lipps, J. & Combellick, R. *Geol. Soc. Am. Bull.* **117**, 996–1008 (2005).
- Shennan, I. *et al. Quat. Sci. Rev.* **15**, 1–37 (1996).
- Shennan, I. *et al. Quat. Sci. Rev.* **17**, 365–393 (1998).
- Shennan, I., Scott, D., Rutherford, J. & Zong, Y. *Quat. Int.* **60**, 55–73 (1999).
- Zong, Y., Shennan, I., Combellick, R. A., Hamilton, S. L. & Rutherford, M. M. *Holocene* **13**, 7–20 (2003).
- Dragert, H., Wang, K. & James, T. S. *Science* **292**, 1525–1528 (2001).
- Miller, M. M., Melbourne, T., Johnson, D. J. & Sumner, W. Q. *Science* **295**, 2423 (2002).

through narrow metal strips. This has proved problematic: when a propagating plasmon is squeezed from the sides, its propagation length is severely reduced². The several strategies deployed to circumvent this problem all represented an imperfect trade-off between sub-wavelength lateral confinement and propagation length.

In short, a new actor was needed: enter the channel plasmon–polariton (CPP). The fundamental idea of guiding light along the bottom of milled V-grooves in a planar metal surface was proposed³ 15 years ago and subsequently refined⁴. Compared with other plasmons, these channel modes are laterally strongly confined and suffer only low propagation loss. Extensive numerical simulations⁵ have confirmed this, and also showed that so-called single-mode operation — meaning that, at a given wavelength, energy (information) is transmitted at one speed — can be attained in such waveguides simply by adjusting the depth of the grooves. (Single-mode operation is advantageous where optical interconnectors are used, because in multi-mode operation, light can jump between modes at a junction, provoking a distortion in the shape of the light pulse.) That finding was followed by the prediction last year⁶ that transmission of light through a sharp 90° bend in a CPP waveguide was possible almost without loss. Waveguides created in dielectric materials with a 'photonic band-gap' (that is, in which light in a certain wavelength range cannot propagate through the bulk structure) could do the same job, but only at the price of a much larger device.

So much for the theory. On the experimental side, too, things have begun to evolve rapidly. Just eight months ago, Bozhevolnyi and colleagues reported⁷ the experimental achievement of CPP propagation along a straight, V-groove waveguide drilled in a gold film using focused ion-beam milling. Working at telecommunication wavelengths between 1.4 and 1.6 μm , they used a near-field optical microscope to build up a picture of the propagation all along the waveguide. They thus showed that the light was confined to a width of around 1 μm that was less than its wavelength, and had a propagation length of 90–250 μm , depending on the exact wavelength.

In their latest contribution¹, Bozhevolnyi *et al.* fabricate the first CPP-based optical components. The first of these is a 'Y-splitter', a junction in which two straight waveguides are connected to a third over a distance of only 5 μm , just over three times the light's wavelength. This feature is of paramount importance for the implementation of miniature optical circuits on a chip.

As a proof of principle, the authors also demonstrate very high performance for a Mach–Zehnder interferometer (in effect, two Y-splitters fork-on-fork that can split and then reunify a light beam) and a functional ring resonator. This latter component (see Fig. 3 on page 510) can — according to the phase

SOLID-STATE PHYSICS

Light at the end of the channel

Francisco J. Garcia-Vidal

If photonic circuits are ever to compete with their electronic counterparts, strong confinement of light waves coupled with low propagation losses is needed. A new class of waveguides offers both.

Miniaturized circuits that use light to carry digital information would be inherently faster than conventional electronic circuits, and have a capacity thousands of times greater. But there's a snag: the development of practical, small photonic components is impeded by the diffraction limit — the fact that light will spread out on passing through any region narrower than its wavelength. On page 508 of this issue¹, Bozhevolnyi *et al.* flag a new route around this obstacle. They present the first components that guide and manipulate light in the form of so-called channel plasmon–polaritons. These guide the light along the bottom of sub-wavelength V-shaped grooves, milled in a metal film, without high propagation losses.

Channel plasmon–polaritons are young members of an extended family known as the surface plasmons. These are electromagnetic waves that originate in the collective excitation of free electrons at the interface of a metal and an insulating dielectric, such as air. Surface plasmons remain tightly bound to the inter-

face: a plasmon of an optical wavelength — between about 400 and 750 nanometres — penetrates around 10 nm into the metal and decays over a few hundred nanometres in the dielectric.

Surface plasmons thus concentrate light in a volume less than its wavelength across. They can also be used to transmit electromagnetic signals: for the near-infrared wavelengths around 1.5 micrometres, typically used in telecommunications, the propagation length of a plasmon at a planar gold–air interface is about a millimetre, and therefore long enough to connect two devices on a chip optically. The use of surface plasmons is also compatible with available planar electronics technology, also offering the possibility of transporting optical signals and electrical current on the same substrate.

But to create miniature photonic circuits, surface plasmons have to be confined not just in the direction perpendicular to the interface, but also in the plane of the interface, so that they can propagate efficiently

difference of the light waves entering and leaving the ring, and therefore the degree of constructive or destructive interference between the two — act as a wavelength filter for the transmitted light.

The fabrication process exploits current planar technology: milling grooves onto a metal film with a focused ion beam is similar to drawing lines on a paper with a pencil. The possibilities for such techniques are huge, but there are still some problems. One is how to get external light into a CPP waveguide and extract it at the other end. Coupling to a standard single-mode optical fibre, as practised by Bozhevolnyi and colleagues^{1,7}, might produce large losses, and other alternatives should be tested. Nevertheless, the successes already achieved in a very short period of experimental research using channel

plasmons to mould the flow of light hint at bright prospects ahead.

Francisco J. Garcia-Vidal is in the Departamento de Física Teórica de la Materia Condensada, Universidad Autónoma de Madrid, Madrid 28049, Spain.
e-mail: fj.garcia@uam.es

1. Bozhevolnyi, S. I., Volkov, V. S., Devaux, E., Laluet, J.-Y. & Ebbesen, T. W. *Nature* **440**, 508–511 (2006).
2. Lamprecht, B. et al. *Appl. Phys. Lett.* **79**, 51–53 (2001).
3. Lu, J. Q. & Maradudin, A. A. *Phys. Rev. B* **42**, 11159–11165 (1990).
4. Novikov, I. V. & Maradudin, A. A. *Phys. Rev. B* **66**, 035403 (2002).
5. Gramotnev, D. K. & Pile, D. F. P. *Appl. Phys. Lett.* **85**, 6323–6325 (2004).
6. Pile, D. F. P. & Gramotnev, D. K. *Opt. Lett.* **30**, 1186–1188 (2005).
7. Bozhevolnyi, S. I., Volkov, V. S., Devaux, E. & Ebbesen, T. W. *Phys. Rev. Lett.* **95**, 046802 (2005).

MATHEMATICAL PHYSICS

Going to ground

Christos N. Likos

How can one find the minimum total energy of an infinite number of particles? A proof showing that, for certain interactions, periodic 'ground states' exist provides a new perspective on this, one of the oldest questions in physics.

When deciding how particles move and interact, nature has an affinity for minimum values. In newtonian mechanics, particles choose trajectories that minimize action, a quantity with the dimensions energy $\times$ time. In thermodynamics, a collection of particles at a fixed volume and temperature will settle into a state that minimizes the so-called free energy of the system, a quantity that, at absolute zero, reduces to the total energy — that is, the sum of all the interactions between pairs of particles.

Simple as such statements may seem, the task of finding 'ground-state' configurations, and proving that they do indeed minimize the energy of a system, is notoriously difficult. The standard approach is more or less trial and error: one chooses a number of configurations, calculates their energies, and picks the structure with the smallest energy as the ground state.

Writing in *Physical Review Letters*¹, András Sütő short-circuits this process. He demonstrates with a strikingly powerful mathematical argument that, in certain cases, the ground-state configuration sought when a material freezes into a regular crystal structure is a periodic one. Certain formal criteria must be met: first, the pair-interaction potential $\varphi(r)$, which describes the potential energy of a two-particle system in terms of their separation, r , must possess a mathematical analogue known as a Fourier transform. (This transform

expresses the potential as a series of oscillating sine and cosine terms that depend on a parameter k , the wavenumber, which represents an inverse wavelength.) Additionally, the Fourier-transformed potential, which is written $\hat{\varphi}(k)$, must be positive for all values of k below some threshold value, K_0 , and zero for all values above it.

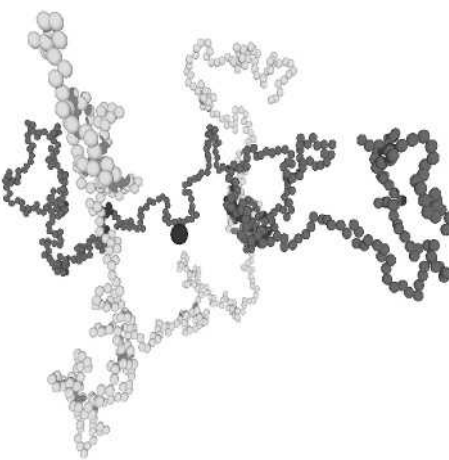


Figure 1 | Chain overlap. A snapshot from a simulation involving two self-avoiding polymers. In this configuration, the centres of mass of the two chains (denoted by the big sphere) can fully overlap. Sütő's results¹, showing that for certain interactions periodic ground states exist, deal with a similar type of interaction that was long considered 'unphysical'. (Courtesy of Arben Jusufi)

The Fourier description of the potential is convenient for potentials such as those considered by Sütő, because it allows the total lattice energy — the energy required to bring all particles infinitely far apart from each other — to be expressed as a simple mathematical formula. The advantage of the Fourier representation is clearly seen when the potential not of particle pairs, but of a periodic lattice of particles, is considered. Such a spatial lattice — a direct, or Bravais, lattice — also has a Fourier representation, known as the reciprocal lattice.

In the reciprocal lattice, the total energy of the system (the quantity we wish to minimize) is simply the sum, up to K_0 , of the lattice potentials $\hat{\varphi}(\mathbf{K})$ for all reciprocal-lattice vectors $\mathbf{K}$ that build up the lattice. (For a three-dimensional reciprocal lattice, each $\mathbf{K}$ will have three wavenumber components, and so uniquely define a point in the lattice.) Whereas the term at $\mathbf{K} = (0, 0, 0)$ represents the origin of the coordinate system and is therefore the same for all types of lattices, all other allowed energy contributions, and thus the total energy of the system, depend on the exact spatial configuration of the particles in the direct lattice.

Imagine now a direct lattice (call it X) for whose reciprocal lattice all non-zero reciprocal-lattice vectors define points outside a sphere of radius K_0 . In this case, the total energy of the system will include only the term at $\mathbf{K} = (0, 0, 0)$. But for all other lattices with at least one value of $\mathbf{K}$ smaller than K_0 , the total energy can only be larger, because the additionally contributing potentials $\hat{\varphi}(\mathbf{K})$ are, by our definition, positive. So X must have the minimum possible energy.

Sütő¹ proved his point by showing that, for a three-dimensional lattice and at the minimum particle density for which his results apply, given by $\rho^* = K_0^3/(8\sqrt{2}\pi^3)$, the body-centred cubic (bcc) lattice that many metals assume indeed fulfils the above requirements. With increasing density, other lattice types join the group, producing an infinite number of ground-state configurations. Moreover, Sütő demonstrated that periodic ground-state configurations are stable against arbitrary deviations from periodicity, and that aperiodic unions of periodic configurations minimize energy.

These are remarkable achievements. But why did such a clear and powerful proof take so long? The answer probably lies in the peculiar form of $\varphi(r)$, which implicitly demands a finite potential at zero separation — the equivalent of permitting two particles to overlap each other fully. In the case of interacting atoms, however, a strong repulsive force is created between bound electron shells and their quantum-mechanical wavefunctions because of the Pauli exclusion principle (which holds that no two particles such as electrons may occupy

CELL DIVISION

Running rings around the spindle

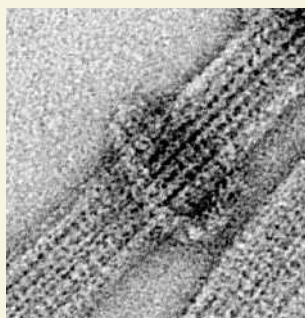
When cells divide, each chromosome is duplicated and one copy is passed to each daughter cell. The details of how these duplicated chromosomes are separated are still being puzzled out, but elsewhere in this issue, Georjana Barnes and colleagues (*Nature* **440**, 565–569; 2006) add a large piece to the jigsaw.

During cell division, an array of tiny filaments called microtubules spreads from the poles of the cell to the central plane, where the duplicated chromosomes congregate. This microtubule structure — the ‘mitotic spindle’ — is responsible for dragging the two copies of the chromosome apart into the two regions that will form the daughter cells. Microtubules are linear polymers of the protein tubulin of diameter 24 nm that grow or shrink as tubulin monomers are added to or lost from their ends. During chromosome segregation,

the microtubules depolymerize, and thus shorten, from the end nearest the central plane of the cell. So, as the spindle shrinks, the attached chromosomes are carried towards the pole.

The chromosomes stick to the depolymerizing ends of the spindle microtubules through specialized protein complexes known as kinetochores. But if these ends are continually dissolving as the microtubules pull back, how do the chromosomes stay attached? The answer lies with a subcomplex in the kinetochore, called the Dam1 complex, which assembles into rings and paired helices that encircle the ends of depolymerizing microtubules (pictured).

Barnes *et al.* used real-time imaging to watch the Dam1 complex on dynamic microtubules *in vitro*. They labelled the Dam1 complex and the microtubules with different coloured fluorophores. As the



microtubules shortened during depolymerization, the Dam1 complex could clearly be seen progressing along the microtubules. This movement could be explained either by a ‘sliding’ mechanism, in which the Dam1 complex is continually associated with the microtubule lattice, or by a ‘turnover’ mechanism, with new subunits of the Dam1 complex rapidly incorporating into existing rings as old subunits leave, thus propelling

the ring along the microtubule. The authors demonstrated that turnover of Dam1-complex subunits was not taking place and that the ring structure is sliding intact along the depolymerizing microtubule.

Finally, Barnes and colleagues showed that the Dam1 complex can function as a coupling device to move objects along the microtubules. These experiments relied on the strong bonds formed between streptavidin and biotin molecules. Streptavidin-coated microbeads were recruited to the microtubules using a biotin-labelled Dam1 complex. As the microtubules depolymerized, the Dam1 complex slid along the spindle away from the shrinking ends, bearing its cargo of beads. It will be interesting to find out whether chromosomes are transported in the same way.

Deepa Nath

G. BARNES *ET AL.*

the same quantum state), and the atomic-pair potential diverges steeply towards infinity as r approaches zero. And at the mesoscopic scale — that between the microscopic and the macroscopic — the rigid core of colloidal particles ensures that they just ‘bounce off’ each other at low separations, and also do not overlap. Neither the atomic nor the mesoscopic potentials even possess a Fourier transform; for a long time, finite interactions of the type considered by Sütő¹ were simply considered ‘unphysical’.

But effective potentials between the centres of mass of entities such as block copolymers², polymer chains or dendrimers can indeed remain finite at zero separation (Fig. 1). Theory³ and simulation⁴ have shown that the effective potential between two chains has a gaussian form, and calculations have shown that its ground-state configuration is face-centred cubic (fcc) at low densities and bcc at high densities^{5,6}. Fourier transformation does not change a gaussian form, so $\hat{\phi}(k)$ is positive in this case. There is, however, no value of K_0 above which $\hat{\phi}(k)$ is zero, so Sütő’s result is not valid here. Nevertheless, his work joins a rapidly growing body of research⁷, spurred on by developments in soft-matter physics, on the properties of bounded interaction potentials in one-component systems, in which the properties of $\hat{\phi}(k)$ are used to obtain information on the ground-state configuration or the topology of the phase diagram.

A potential $\phi(k)$ that vanishes above a

threshold seems difficult to realize physically. Sütő shows that the requirement for this to occur leads to an interaction potential that shows oscillatory behaviour combined with a power-law decay at large values of r . This is strikingly similar to the RKKY (Ruderman–Kittel–Kasuya–Yosida), or Friedel, potential^{8–10} that arises as an effective interaction potential between ions when the effect of free electrons is incorporated as a statistical average. For metals of valency Z , the ion density ρ is related to Sütő’s minimum density by $\rho = \rho^* \pi \sqrt{2}/(3Z)$, so that ρ exceeds ρ^* by 48% for $Z=1$ and trails it by 26% for $Z=2$. So unfortunately there is no happy agreement between the real and the Sütő densities that would hint at why certain metals crystallize into the bcc structure. The power laws governing the behaviour of the RKKY and Sütő potentials at large values of r are slightly different, and effective ion potentials also include a steeply repulsive region at small values of r (ref. 10) that is absent in Sütő’s interaction. So the long- and short-range parts of the potential must be considered independently. Sütő provides us with a suitable reference frame to deal with the former.

In the late 1970s, a celebrated and controversial article¹¹ was published entitled ‘Should all crystals be bcc?’. It contained a theory of freezing that involved density waves — modulated density profiles — at the set of reciprocal lattice vectors $\mathbf{K}$ as parameters in a mathematical expression for the free energy. On the basis

that the bcc crystal is the only spatial configuration that combines the shortest reciprocal-lattice vectors lying on the surface of a sphere, it was concluded that the stablest crystals should all be bcc. This argument, although brilliant, was flawed in making certain simplifying assumptions. Sütő’s work comes from a very different direction to again put the bcc lattice in a distinctive position among all Bravais lattices. In this way, it recasts the question of why crystals form as they do in a fresh and thought-provoking way. ■

Christos N. Likos is at the Institut für Theoretische Physik II, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, D-40225 Düsseldorf, Germany.
e-mail: likos@thphy.uni-duesseldorf.de

1. Sütő, A. *Phys. Rev. Lett.* **95**, 265501 (2005).
2. Pierleoni, C., Addison, C., Hansen, J.-P. & Krakoviack, V. *Phys. Rev. Lett.* (in the press); preprint available at www.arxiv.org/abs/cond-mat/0601417 (2006).
3. Krüger, B., Schäfer, L. & Baumgärtner, A. *J. Phys. (Paris)* **50**, 3191 (1989).
4. Louis, A. A., Bolhuis, P. G., Hansen, J.-P. & Meijer, E. J. *Phys. Rev. Lett.* **85**, 2522–2525 (2000).
5. Stillinger, F. H. *J. Chem. Phys.* **65**, 3968–3974 (1976).
6. Stillinger, F. H. *Phys. Rev. B* **20**, 299–302 (1979).
7. Mladek, B. M., Gottwald, D., Kahl, G., Neumann, M. & Likos, C. N. *Phys. Rev. Lett.* **96**, 045701 (2006).
8. Ashcroft, N. W. & Mermin, N. D. *Solid State Physics* (Holt Saunders, Philadelphia, 1976).
9. Ziman, J. M. *Principles of the Theory of Solids* (Cambridge Univ. Press, 1972).
10. Hafner, J. *From Hamiltonians to Phase Diagrams* (Springer, Berlin, 1987).
11. Alexander, S. & McTague, J. *Phys. Rev. Lett.* **41**, 702–705 (1978).

BRIEF COMMUNICATIONS

Influenza virus receptors in the human airway

Avian and human flu viruses seem to target different regions of a patient's respiratory tract.

Although more than 100 people have been infected by H5N1 influenza A viruses, human-to-human transmission is rare¹. What are the molecular barriers limiting human-to-human transmission? Here we demonstrate an anatomical difference in the distribution in the human airway of the different binding molecules preferred by the avian and human influenza viruses. The respective molecules are sialic acid linked to galactose by an α -2,3 linkage (SA α 2,3Gal) and by an α -2,6 linkage (SA α 2,6Gal)². Our findings may provide a rational explanation for why H5N1 viruses at present rarely infect and spread between humans although they can replicate efficiently in the lungs.

SA α 2,3Gal molecules have been found on cells artificially differentiated from isolated human tracheal and bronchial cells *in vitro*³. But the anatomical distribution and prevalence of SA α 2,3Gal and SA α 2,6Gal in the human airway was unknown. Using lectins specific for SA α 2,3Gal and SA α 2,6Gal, we found that SA α 2,6Gal is dominant on epithelial cells in nasal mucosa, with SA α 2,3Gal being occasionally detected (for details and methods, see supplementary information). Epithelial cells in the paranasal sinuses, the pharynx, the trachea⁴ and the bronchi mainly express SA α 2,6Gal (Fig. 1, and see supplementary information). In the terminal and respiratory bronchioles, epithelial cells also express SA α 2,6Gal sialyloligosaccharides.

SA α 2,3Gal was found on non-ciliated cuboidal bronchiolar cells at the junction between the respiratory bronchiole and alveolus, however, and a substantial number of cells lining the alveolar wall also expressed this molecule. The SA α 2,3Gal-positive alveolar cells also reacted to an antibody against surfactant protein A; this suggests that they were alveolar type-II cells (which express surfactant protein A). Alveolar type-II cells have been shown to be infected with H5N1 virus in a patient⁵.

Human-derived viruses that preferentially recognize SA α 2,6Gal bound extensively to epithelial cells in the bronchi and, to a lesser degree, to alveolar cells; by contrast, avian viruses that preferentially recognize SA α 2,3Gal bound extensively to alveolar cells but less widely to bronchial epithelial cells (see supplementary information). It is interesting that the A/Hong Kong/213/03 (H5N1) virus, which was isolated from a human and recognizes both SA α 2,3Gal and SA α 2,6Gal (ref. 6), bound

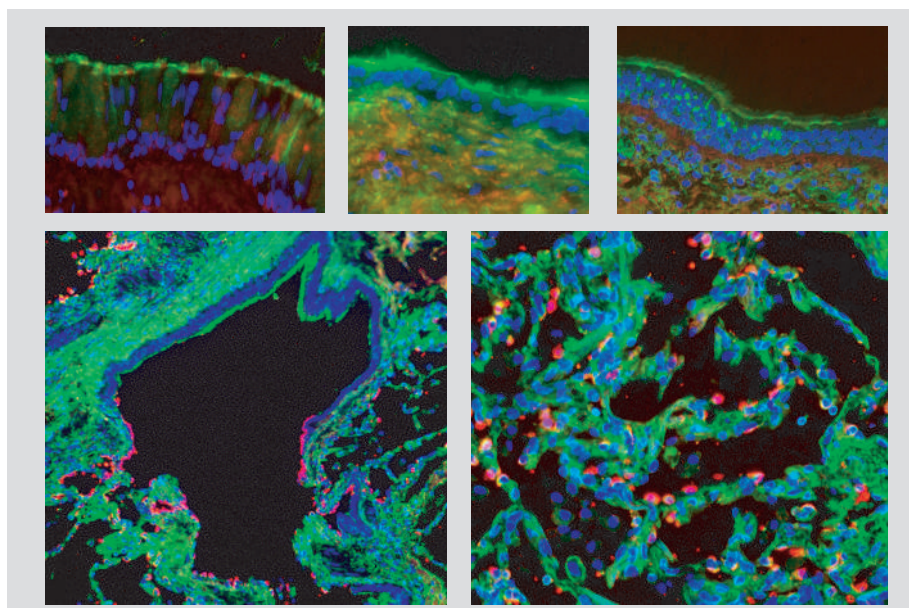


Figure 1 | Reactivity of human respiratory tissues with lectins specific for different sialic acid linkages. **a**, Nasal mucosa; **b**, paranasal sinuses; **c**, bronchus; **d**, bronchiole; **e**, alveolus. Res, respiratory bronchiole (adjacent to alveoli); Ter, terminal bronchiole (distal to alveoli); Alv, alveolus. Green, reaction with *Sambucus nigra* lectin, indicating the presence of sialic acid linked to galactose by an α 2,6-linkage (SA α 2,6Gal). Red, reaction with *Maackia amurensis* lectin, indicating the presence of SA α 2,3Gal. Cells were counterstained with DAPI (4,6-diamidino-2-phenylindole).

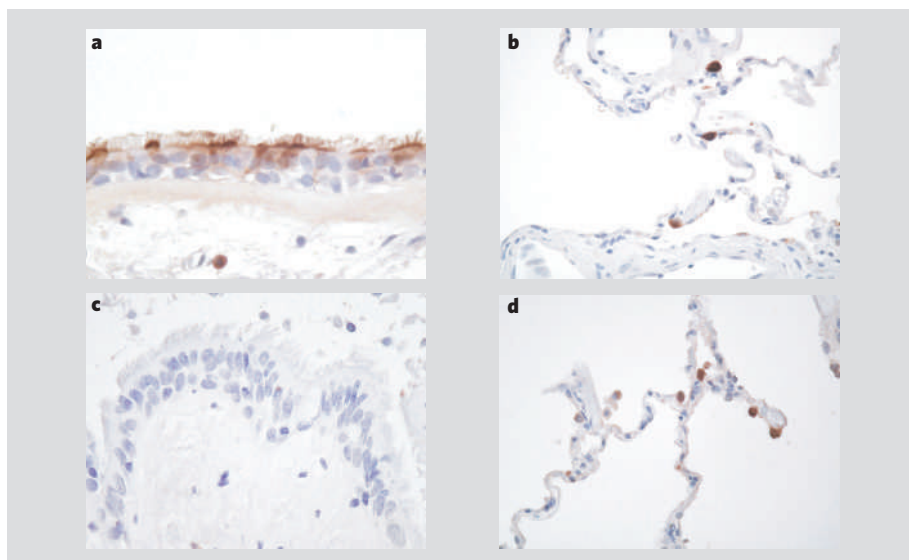


Figure 2 | Infection of human respiratory tissue by influenza A viruses. **a**, Extensive infection of bronchial epithelial cells by human A/Kawasaki/173/01 (H1N1) virus. **b**, Infection of alveolar cells by the human virus. **c**, Bronchial epithelial cells are not infected by A/duck/Mongolia/301/01 (H3N2) virus. **d**, Infection of alveolar cells by the avian virus. The results obtained with other human viruses, A/Yokohama/2057/03 (H3N2) and A/Kawasaki/176/02 (H1N1), were similar to those obtained with A/Kawasaki/173/01, and those obtained with the avian viruses A/duck/Czechoslovakia/56 (H4N6) and A/duck/Vietnam/5001/05 (H5N1) were similar to those obtained with A/duck/Mongolia/301/01. Infected cells are stained brown.

extensively to both bronchial and alveolar cells.

Human-derived viruses preferentially recognizing SA α 2,6Gal efficiently infected epithelial cells lining the bronchi and alveolar cells (Fig. 2a, b), whereas avian viruses preferentially recognizing SA α 2,3Gal infected alveolar cells but not bronchial epithelial cells (Fig. 2c, d, and see supplementary information). As shown in a virus-binding assay, A/Hong Kong/213/03 infected both alveolar cells and epithelial cells in bronchi (see supplementary information). Although not quantitative, these results indicate that there could be functional significance in the preferential expression of SA α 2,3Gal and SA α 2,6Gal molecules on human airway cells.

Our findings indicate that although H5N1 viruses preferentially recognizing SA α 2,3Gal can be transmitted from birds to humans, they can replicate efficiently only in cells in the lower region of the respiratory tract, where the avian-virus receptor is prevalent. This restriction may contribute to the inefficient human-to-human transmission of H5N1 viruses seen so far, and indicates that unimpeded transmission of the virus might require acquisition of the ability

to recognize SA α 2,6Gal. This change would enable the virus to replicate in the upper region of the respiratory tract, where it could be readily spread by sneezing and coughing. Mutations in the viral haemagglutinin molecule would be necessary to confer SA α 2,6Gal-binding ability on H5N1 avian viruses, although amino-acid substitutions in other viral proteins, including PB2 (refs 7, 8), may be required to confer pandemic potential on avian viruses that can efficiently replicate in humans.

Kyoko Shinya*†‡§, **Masahito Ebina**||, **Shinya Yamada**†, **Masao Ono**¶, **Noriyuki Kasai**‡, **Yoshihiro Kawaoka***†#☆

*Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA
e-mail: kawaoka@svm.vetmed.wisc.edu

†Division of Virology, Department of Microbiology and Immunology, #International Research Center for Infectious Diseases, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan

‡Institute for Animal Experimentation, ||Respiratory Oncology and Molecular Medicine, Institute of Development, Aging, and Cancer,

¶Department of Pathology, Graduate School of Medicine, Tohoku University, Sendai, Miyagi 980-8575, Japan

☆Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan

§Present address: The Avian Zoonosis Research Centre, Tottori University, Tottori 680-8550, Japan

1. Beigel, J. H. *et al.* *N. Engl. J. Med.* **353**, 1374–1385 (2005).
2. Rogers, G. N. & Paulson, J. C. *Virology* **127**, 361–373 (1983).
3. Matrosovich, M. N., Matrosovich, T. Y., Gray, T., Roberts, N. A. & Klenk, H. D. *Proc. Natl Acad. Sci. USA* **101**, 4620–4624 (2004).
4. Baum, L. G. & Paulson, J. C. *Acta Histochem.* **40** (suppl.), 35–38 (1990).
5. Uiprasertkul, M. *et al.* *Emerg. Infect. Dis.* **11**, 1036–1041 (2005).
6. Shinya, K. *et al.* *J. Virol.* **79**, 9926–9932 (2005).
7. Naffakh, N., Massin, P., Escriou, N., Crescenzo-Chaigne, B. & van der Werf, S. *J. Gen. Virol.* **81**, 1283–1291 (2000).
8. Hatta, M., Gao, G., Halfmann, P. & Kawaoka, Y. *Science* **293**, 1840–1842 (2001).

Supplementary information accompanies this communication on Nature's website.

Received 12 Feb 2006; accepted 9 March 2006.

Competing financial interests: declared none.

doi:10.1038/440435a

ECOLOGY

Human role in Russian wild fires

Anomalies in temperature and precipitation in northern Russia over the past few years have been viewed as manifestations of anthropogenic climate change¹, prompting suggestions that this may also account for exceptional forest fires in the region^{2,3}. Here we examine the number of forest-fire events across the boreal Russian Federation for the period 2002 to 2005 in 'intact' forests, where human influence is limited, and in 'non-intact'

forests, which have been shaped by human activity⁴. Our results show that there were more fires in years during which the weather was anomalous, but that more than 87% of fires in boreal Russia were started by people.

Since the last Ice Age, fire has been the main process that has disturbed the physical features of the boreal biome. Three major fire seasons have occurred recently in Russia: a total of 6.9, 7.5 and 14.5 million hectares (ha) were burned

in 1998 (ref. 5), in 2002 (ref. 5) and in 2003 (ref. 6), respectively. Large climate anomalies occurred during this period⁷ that increased the likelihood of fire ignition and propagation. In combination, this evidence points to a serious impact of climate change on Russian forests^{2,3}, although fires are considered to be caused predominantly by humans in 'normal' years^{4,8}.

Large areas of intact forest still exist in Russia; it can be assumed that human influence on these areas is limited. Using satellite data to locate fires, we estimated the number of fire events in these intact forests and in forests more affected by humans in order to assess the specific impact of climate anomalies

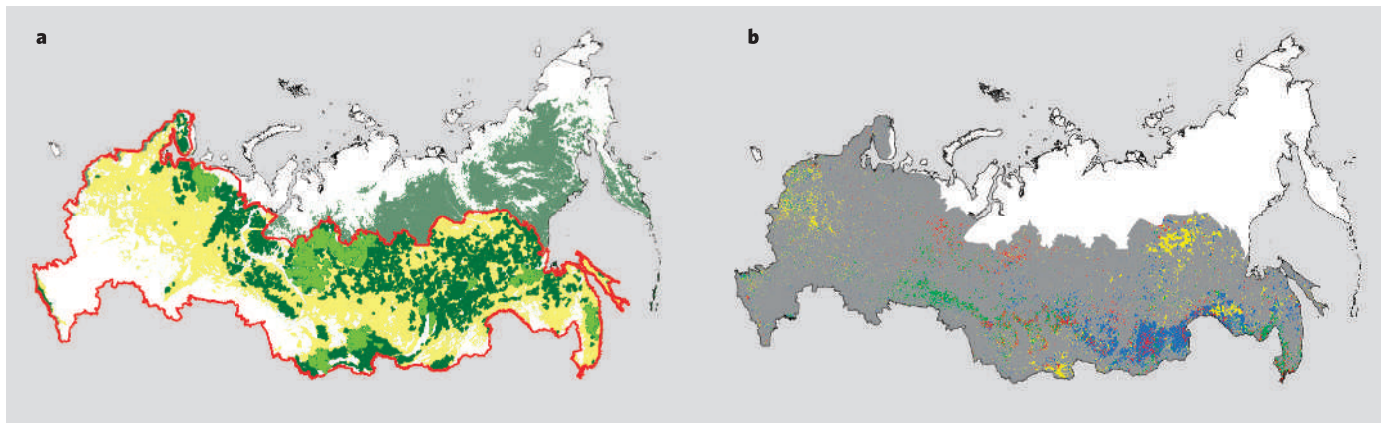
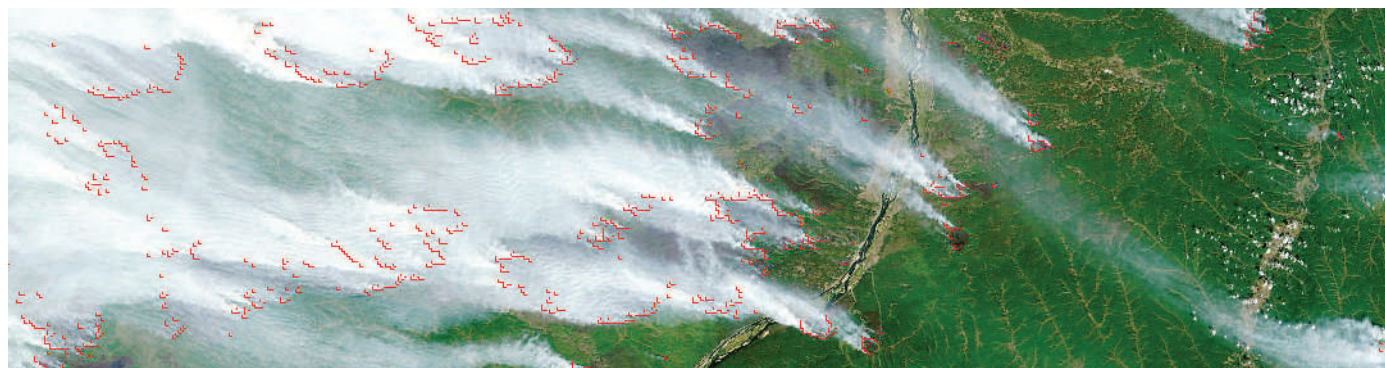


Figure 1 | Maps of the Russian Federation showing areas of intact forest and the locations of forest fires over 2002–2005. a, The 'intact' forest areas (dark green) of the region are defined⁴ as being situated within the forest zone, having a size exceeding 50,000 ha and a width of 10 km at the narrowest point. They contain a contiguous mosaic of natural ecosystems, are not fragmented by infrastructure, show no signs of significant transformation by humans, and have a natural fire regime. Light green areas, 'most-intact' forest (a subset of intact forest; see text); yellow, non-intact forest; grey-green, forest outside the study area. The study area is outlined in red. The northern limit for the region covered by the atlas⁴ excludes all tundra ecosystems; the Kamchatka region was excluded from our study zone. **b**, Locations of active fires in 2002 (yellow), 2003 (blue), 2004 (green) and 2005 (red); 2002 and 2003 were severe years. Fire dots are overlaid for years from 2002 to 2005, so fires from earlier years may be obscured by later years (dot size is exaggerated for display purposes). The study area is shaded in grey.



MODIS RAPID RESPONSE PROJECT AT NASA/GSFC

Smoking gun: an image at 1-km resolution from the Terra satellite in August 2002 shows forest fires (red polygons) with white smoke plumes.

(inside intact forests) and the combined impact of humans and climate change (outside intact forests).

We combined an atlas of Russia's intact forest landscapes⁴ with a land-cover map⁹ drawn up for 1999–2000 to delineate intact and non-intact forests (see supplementary information). The total area covered by the atlas⁴ is 1,118 million ha, including 205 million ha of intact forests. Active fires detected by the Terra satellite¹⁰ were used to derive our fire database for the period 2002 to 2005 (Fig. 1). These active fires were separated into those that fell inside or outside intact forests. Active fires that were spatially and temporally concurrent were then grouped together to identify individual fire events. Finally, fire events were attributed to the zone in which they started.

We found that 85% of the 23,818 active fires in intact forests in 2002 were located within a 10-km-wide buffer zone inside the perimeter of the intact forests, indicating that the borders of these areas could have been affected by fires starting on the outside. We therefore selected a new subset of 'most-intact' forests to represent more strictly the concept of intact forest — these comprised the largest individual intact areas and the individual intact areas with the smallest perimeter-to-area ratio. The

resulting subset of 17 intact areas totals 25% of the total intact area.

During seasons of climatic anomalies (2002 and 2003), fire-event densities in the most-intact forests were twice the normal density (Table 1). But, more surprisingly, density ratios of forest-fire events outside and inside most-intact forests are at least 7.9 and as high as 14.4 during the four-year study period. These results show that human impact had a constant 'multiplication' effect on the fire events, and that a maximum of 13% of fire ignition in the non-intact forests is explained by natural disturbance, with the rest being directly induced by humans.

Forest regions of boreal Eurasia, and particularly Siberia, have seen a reduction in population following the creation of the Russian Federation. But the human impact on the forests through fires is higher owing to lack of control, ineffectual fire-management policies and new socioeconomic conditions in the region^{7,11,12}. It is also a consequence of the oil boom in Siberia¹³. The fact that recent increases in 'wild' fires in Eurasian boreal forests are primarily a result of human behaviour on the ground has implications for the global carbon budget³ and should be taken into account in future mitigation policies.

Table 1 | Number and density of fire events in forests of the Russian Federation

	2002	2003	2004	2005
Forest type*	Number of fire events			
All forests in study region	5,929	6,461	4,627	3,796
Non-intact forests	5,420	5,967	4,417	3,477
Intact forests	509	494	210	319
Most-intact forests	109	96	49	119†
	Density of fire events (10 ⁻³ per km ²)			
All forests in study region	1.09	1.19	0.85	0.70
Non-intact forests	1.60	1.77	1.31	1.03
Intact forests	0.248	0.241	0.102	0.128
Most-intact forests	0.202	0.178	0.091	0.115
	Density ratio of fire events			
Ratio non-intact/intact	6.5	7.3	12.8	8.1
Ratio non-intact/most-intact	7.9	9.9	14.4	9.0

*The area of all forests in the study zone is 543 million ha, of which 338 million ha is non-intact forest and 205 million ha is intact forest. The total area of 'most-intact' forest (a subset of intact forest; see text) is 54 million ha. The fire season is considered to last until 21 September.

†From these 119 fire events in most-intact forests, 57 are located in two large intact forest areas in the basin of the Taz River, where oil was being prospected (as revealed by fine spatial resolution imagery) in 2005. Intact areas were delineated using satellite imagery from the year 2000; some areas may no longer have been intact in 2005, in particular as a result of the oil boom in Siberia¹³. Fire-event densities are estimated without these 57 fires.

Danilo Mollicone, Hugh D. Eva, Frédéric Achard

Institute for Environment and Sustainability,
Joint Research Centre of the European
Commission, TP 440, 21020 Ispra, Italy
e-mail: frederic.achard@jrc.it

1. Thompson, D. W. J. & Wallace, J. M. *Science* **293**, 85–89 (2001).
2. Dale, V. H. *et al. BioScience* **51**, 723–734 (2001).
3. Schiermeier, Q. *Nature* **435**, 732–733 (2005).
4. Aksenov, D. *et al. Atlas of Russia's Intact Forest Landscapes* (Global Forest Watch Russia, Moscow, 2002); available at <http://forest.ru/eng/publications/intact/>
5. Sukhinin, A. *et al. Remote Sens. Environ.* **93**, 546–564 (2004).
6. Goldammer, J. G., Sukhinin, A. & Csizsar, I. *Int. Forest Fire News* [online] **29**, <http://www.fire.uni-freiburg.de/iffn/iffn_online.htm> (2003).
7. International Research Institute for Climate Prediction IRI Map Room (2005); available at <http://iridl.ldeo.columbia.edu/maproom>
8. Mouillot, F. & Field, C. B. *Global Change Biol.* **11**, 398–420 (2005).
9. Bartalev, S., Belward, A. S., Erchov, D. & Isaev, A. S. *Int. J. Remote Sens.* **24**, 1977–1982 (2003).
10. Justice, C. O. *et al. Remote Sens. Environ.* **83**, 244–262 (2002).
11. Achard, F., Stibig, H.-J., Laestadius, L., Yaroshenko, A. & Aksenov, D. *Identification of "Hot Spot Areas" of Forest Cover Changes in Boreal Eurasia* (European Commission, Luxembourg, 2005); available at <http://ies.jrc.ec.eu.int/364.html>
12. Karpachevskiy, M. *Forest Fires in the Russian Taiga: Natural Disaster or Poor Management?* (Taiga Rescue Network Factsheet, Jokkmokk, Sweden, 2004).
13. Dienes, L. *Eurasian Geogr. Econ.* **45**, 319–345 (2004).

Supplementary information accompanies this communication on Nature's website.

Received 11 November 2005; accepted 2 March 2006.

Competing financial interests: declared none.

doi:10.1038/440436a

RETRACTION

Cell biology: Non-thermal heat-shock response to microwaves

D. de Pomerai, C. Daniells, H. David, J. Allan, I. Duce, M. Mutwakil, D. Thomas, P. Sewell, J. Tattersall, D. Jones, P. Candido
Nature **405**, 417–418 (2000)

Our claim that weak microwave fields induce a heat-shock response in *Caenorhabditis elegans* by a non-thermal mechanism is invalidated by new findings showing that there is a small heating effect under these conditions (A. Dawe *et al. Bioelectromagnetics* **27**, 88–97; 2006). This temperature rise (about 0.2 °C) causes heat-shock induction comparable to that noted in our communication.

C.D., J.A., M.M., D.J. and P.C. were not available to sign this retraction.

doi:10.1038/440437a

**Cover illustration**

View looking into a nicotinic acetylcholine receptor from *Torpedo marmorata*, courtesy of Nigel Unwin, MRC Laboratory of Molecular Biology.

Editor, Nature

Philip Campbell

Insights Publisher

Sarah Greaves

Insights Editor

Lesley Anson

Production Editor

Maria Hodges

Senior Art Editor

Martin Harrison

Art Editor

Nik Spencer

Layouts

Jane Walker

Sponsorship

Claire Hines

Claudia Banks

Production

Jocelyn Hilton

Marketing

Claire Aspinall

Editorial Assistant

Laura Shaw

ION CHANNELS

What are ion channels? And why are so many people excited about them at the moment? In a nutshell, ion channels are fundamental to cellular existence. They are present in the membranes of almost all living cells — from simple bacteria to highly specialized neurons — and are essential for important physiological processes such as sensory transduction, action-potential generation and muscle contraction, to name just a few. Indeed, when their function goes awry, there can be serious consequences, including life-threatening disease.

Ion channels are proteins that allow the passage of charged ions through hydrophobic membranes. But they are not simple holes; ion channels are sophisticated machines that can conduct ions with exquisite specificity, at speeds close to the limit of diffusion, under very tight regulation. And it is these processes, which until recently have remained largely inaccessible except to some inspired electrophysiological experimentation, that are now being exposed in all their glory by pioneering high-resolution structural studies.

Have these structural insights helped us to derive new treatments for debilitating conditions such as epilepsy, diabetes and cystic fibrosis? Not yet — it is early days for rational drug design — but there are high hopes, particularly as many successful drugs already target ion channels. What is also clear is the extent to which research on disease genes has contributed to our understanding of the normal physiological roles that ion channels have and their underlying mechanisms — mechanisms that will continue to fascinate and perplex ‘channelologists’ for years to come.

We are most grateful to the many people who contributed to this Insight, whether by writing, advising or generating the stimulating data discussed in the following pages.

Lesley Anson, Senior Editor

REVIEWS

- 440 From molecule to malady**
F. Ashcroft
- 448 Recent advances in Cys-loop receptor structure and function**
S. M. Sine & A. G. Engel
- 456 Glutamate receptors at atomic resolution**
M. L. Mayer
- 463 hERG potassium channels and cardiac arrhythmia**
M. C. Sanguinetti & M. Tristani-Firouzi
- 470 K_{ATP} channels as molecular sensors of cellular metabolism**
C. G. Nichols
- 477 The ABC protein turned chloride channel whose failure causes cystic fibrosis**
D. C. Gadsby, P. Vergani & L. Csanády
- 484 CIC chloride channels viewed through a transporter lens**
C. Miller

nature
insight

From molecule to malady

Frances M. Ashcroft¹

Ion channels are membrane proteins, found in virtually all cells, that are of crucial physiological importance. In the past decade, an explosion in the number of crystal structures of ion channels has led to a marked increase in our understanding of how ion channels open and close, and select between permeant ions. There has been a parallel advance in research on channelopathies (diseases resulting from impaired channel function), and mutations in over 60 ion-channel genes are now known to cause human disease. Characterization of their functional consequences has afforded unprecedented and unexpected insights into ion-channel mechanisms and physiological roles.

The cell membrane is a major barrier to ion movement, and specific proteins — the ion channels, transporters and pumps — have evolved to transport ions across it. Ion channels are gated pores that permit the passive flow of ions down their electrochemical gradients (Fig. 1). Ion pumps, by contrast, use the energy of ATP hydrolysis to transport ions against their electrochemical gradient. In between are the coupled transporters (antiporters and symporters), in which movement of one ion species against its electrochemical gradient is powered by the downhill movement of another.

This Insight focuses on ion channels. Over 340 human genes are thought to encode ion channels (see www.ensembl.org/Homo_sapiens/index.html). They have important roles in such diverse processes as nerve and muscle excitation, hormone secretion, cell proliferation, sensory transduction, learning and memory, regulation of blood pressure, salt and water balance, lymphocyte proliferation, fertilization and cell death¹. Your ability to read and understand this page depends on the activity of ion channels in your eye and brain. Consequently, defects in ion-channel function often have profound physiological effects.

Because of their important functional roles, their membrane location, structural heterogeneity and the restricted tissue expression of some channel types, ion channels are attractive targets for drug therapy. Indeed, many existing drugs, such as local anaesthetics², sedatives³, anti-anxiety agents³, antidiabetic drugs (see the review in this issue by Nichols, p. 470) and even antiviral therapies^{1,4}, exert their effects by interacting with ion channels.

Our understanding of how ion channels function has been illuminated

by recent breakthroughs in high-resolution structure determination and by studies of diseases that result from impaired channel function (the channelopathies). This review provides an introduction to ion-channel structure, a theme that is developed more fully in the subsequent reviews. It also provides an overview of channelopathies.

Structural considerations

Ion channels are composed of one or more pore-forming subunits, often in association with accessory subunits. Most channels conform to a common structural theme in which the central pore, through which the ions move, is formed by four or five transmembrane α -helices, which fit together like the staves of a barrel. In many channels, the pore-forming helices are contributed by separate subunits, so that the channel is tetrameric (Kir channels, for instance) or pentameric (Cys-loop receptors, Fig. 2). However, voltage-gated Ca^{2+} (Ca_v) and Na^+ (Na_v) channels are composed of a single subunit that contains four similar repeated domains, and some K^+ channels are dimers, each subunit being composed of two repeated domains (Fig. 2).

The largest class of ion channels possess a pore loop, a region of the protein that loops back into the membrane to form the selectivity filter that determines which ion species can permeate (Fig. 2). They are either tetramers, or monomers of four similar domains. The human ether-a-go-go (hERG) channels, inwardly rectifying K^+ (Kir) channels and glutamate receptor channels discussed in subsequent reviews belong to this class. All pore-loop channels are built on a common pattern, which suggests that they evolved from a primordial channel resembling

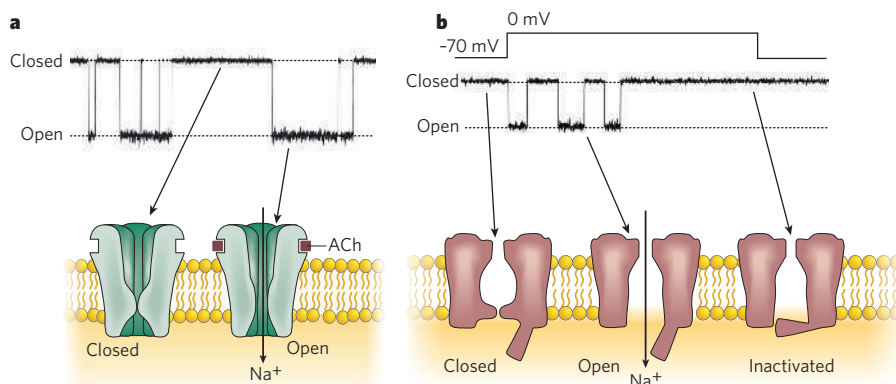


Figure 1 | Molecular nanoswitches. Schematics illustrating how ion channels open and close, with associated single-channel recordings. Opening and closing of the channel are random events, but the frequency with which they occur is influenced by, for example, ligand-binding (a) or transmembrane voltage (b). The transition rate between open and closed states is $<10 \mu\text{s}$. The flux rate through the pore when it is open is of the order of 10^7 ions per second; that mediated by the coupled exchangers is substantially smaller (see p. 484). Following opening, some voltage-gated channels enter an inactivated (non-conducting) state in which they are refractory to subsequent depolarization (b).

¹University Laboratory of Physiology, Parks Road, Oxford OX1 3PT, UK.

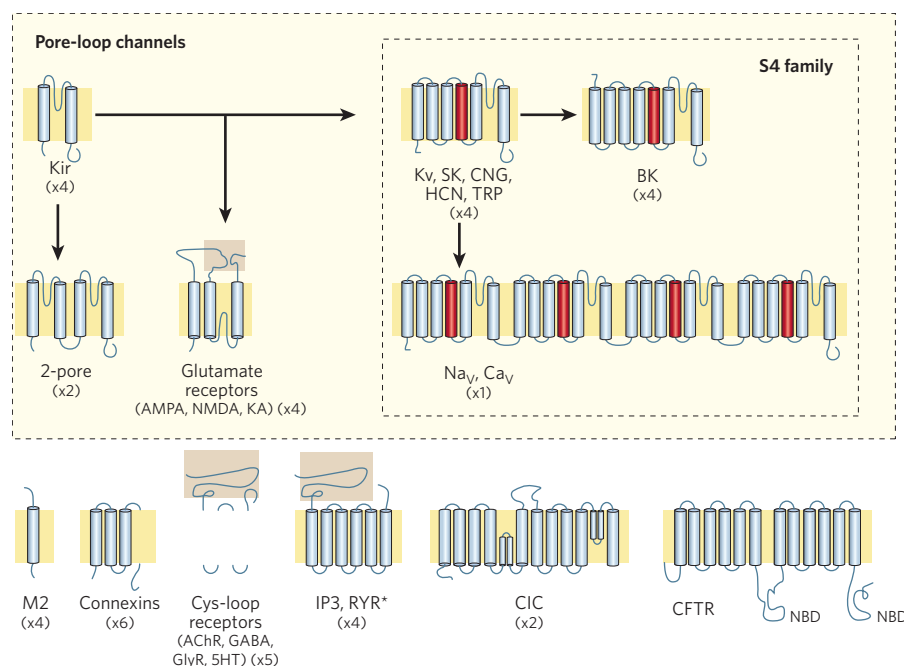


Figure 2 | Ion-channel classes. Predicted transmembrane topology of different types of channel. The numbers in parentheses refer to the number of subunits that make up the channel pore. The beige boxes indicate ligand-binding domains. The S4 domain is highlighted. Kir, inwardly rectifying K⁺ channel. K_v, voltage-gated K⁺ channel. Na_v, voltage-gated Na⁺ channel. Ca_v, voltage-gated Ca²⁺ channel. CNG, cyclic-nucleotide-gated channel. HCN, hyperpolarization-activated channel. TRP, transient receptor potential channel. CIC, chloride channel. CFTR, cystic fibrosis transmembrane conductance regulator. IP3R, inositol trisphosphate receptor. RYR, ryanodine receptor. *The number of TMs of RYRs, whether they are evolutionarily related to IP3R, or if RYR and IP3R belong to the pore-loop family, or not, is unknown. The CIC channels have many helices that go only partway across the membrane and their transmembrane topology is not easily depicted (see p. 484).

the bacterial K⁺ channels KcsA and KirBac^{5,6}. The central core of these channels, which contains the ion-permeation pore, comprises two transmembrane helices (TMs) linked by a pore loop (Fig. 2). The Kir channels consist only of this core, and two-pore K⁺ channels arose by tandem duplication. In the S4 subfamily, additional TMs (mostly four, but five in the case of the large-conductance Ca²⁺-activated K⁺ (BK) channels) are present at the amino-terminal end of the protein⁷. The fourth TM (S4) is adapted for voltage sensing and possesses several positively charged residues. Evolution has further developed the pore-loop theme by addition of extracellular and/or intracellular modules that act as docking sites for ligands, accessory subunits and other modulators. The glutamate receptors, for example, contain a pore-loop core, orientated in the membrane in the inverse direction, plus extracellular ligand-binding domains. Some K⁺ channels have an added intracellular Ca²⁺-binding site⁸, and both cyclic-nucleotide-gated (CNG) channels⁹ and hERG (see the review in this issue by Sanguinetti and Tristani-Firouzi, p. 463) have appropriated an intracellular cyclic-nucleotide-binding site. Like many proteins, therefore, ion channels are built on a modular theme. The tetrameric nature of most pore-loop channels contributes to ion-channel diversity, as closely related subunits can associate to form heteromeric channels with novel properties.

The non-pore-loop channels can be grouped into a variety of classes, which have no obvious evolutionary relationship (Fig. 2). The pentameric ligand-gated channels known as the Cys-loop receptors contain four TMs, the pore being formed by the second transmembrane helix in each of the five subunits. The neurotransmitter-binding site is found in the N terminus and, at least in some cases, may have been stolen from elsewhere. That of the acetylcholine receptor (AChR), for example, is also present in a separate water-soluble protein¹⁰. Interestingly, the Cys-loop receptors are found almost exclusively in metazoans (multicellular organisms), suggesting that they evolved only when cell–cell interactions such as those found in the nervous system became important¹¹.

Also discussed in this Insight are CFTR (see the review in this issue by Gadsby, Vergani and Csanády, p. 477) and the sulphonylurea receptor (SUR; see p. 470). These belong to the ATP-binding cassette (ABC) family of pumps that transport nutrients, drugs and solutes across the membrane using the energy of ATP hydrolysis. Uniquely, CFTR is a Cl[−] channel, and its ATPase activity is used to drive the protein between open and closed conformations (see p. 477). The ABC protein SUR is neither a pump nor a channel but a channel regulator (see p. 470).

The CIC chloride channel family presents a conundrum, for it is now

clear that some of these proteins are channels whereas others are transporters (see the review in this issue by Miller, p. 484). This example, along with that of CFTR and SUR (which ‘should’ be pumps), demonstrates that the classical distinction between channels and transporters is blurred and that the two types of ion-transport mechanism share common features. Understanding why certain family members function as transporters and others as channels will provide fundamental insights into the essential features that distinguish channels from transporters.

High-resolution structures of several ion channels, or their bacterial homologues, have been obtained, including those of Kir⁶ and K_v⁷ channels, the AChR Cys-loop receptor¹³ (see the review in this issue by Sine and Engel, p. 448) and the Cl[−] transporter CIC-ec1¹² (see also p. 484). Although the complete structure of CFTR has not yet been solved, the crystal structure of the first nucleotide-binding domain (involved in ATP hydrolysis) is known (see p. 477). High-resolution structures of the ligand-binding domains of glutamate receptors are also available (see the review in this issue by Mayer, p. 456). These structures, described in detail in the accompanying reviews, have illuminated our understanding of how ion channels work, as discussed below.

Properties of ion channels

The cardinal properties of ion channels are ion selectivity and gating. Selectivity refers to the ability of some channels to discriminate between ion species, allowing some to pass through the pore while excluding others. Gating is the process of transition between the open and closed states (Fig. 1).

Selectivity

Ion movement through a channel pore depends on the electrochemical gradient across the membrane and the permeability of the pore¹⁴. The largest pores, such as those of gap-junction channels^{15,16}, function like molecular sieves and discriminate only on the basis of size. In most cases, however, ion channels are very choosy about the ions they allow through. Some, such as AChR, permit flux of a variety of cations but exclude anions (see p. 448). They have a large water-filled pore (~6.5 Å in diameter) through which ions move in the (partially) hydrated state¹³. Rings of negatively charged residues in the pore exclude anions and facilitate cation flux. Conversely, anion-selective channels such as GABA and glycine receptors possess rings of positively charged residues.

Other channels, particularly those that belong to the pore-loop family, are even more discriminatory. Indeed, a re-entrant pore loop seems

to be a criterion for high selectivity as it is present in all highly selective channels (although the converse is not true — not all pore-loop channels are highly selective between ions of similar charge). K^+ channels are 100–1,000 times more permeable to K^+ than Na^+ , despite the fact that the unhydrated Na^+ ion is smaller¹⁴. This exquisite discrimination takes place at the narrowest part of the pore, known as the selectivity filter. How this is achieved became crystal clear in 1998 when the structure of the bacterial K^+ channel KcsA was solved⁵. All K^+ channels possess an almost invariant GYG sequence that earlier functional studies had shown lies at the heart of the selectivity filter¹⁷. The X-ray structure revealed that the backbone carbonyls of the GYG residues line the pore and substitute for the waters of hydration, so that K^+ passes through the filter by slipping from one binding site to the next⁵. The carbonyl oxygens are too far apart to enable them to interact intimately with Na^+ ions, so that Na^+ is effectively excluded from the selectivity filter because

the energy required to remove the waters of hydration is greater than that gained by interacting with the carbonyl oxygens.

Na_v and Ca_v channels probably use a different strategy to achieve selectivity. Although no structures are yet available, functional studies show that these pores are wider than that of K^+ channels (4–6 Å compared with 3 Å) and that the side chains face into the pore^{14,18,19}. Both types of channel use negatively charged residues, and perhaps some backbone carbonyls as well, to coordinate the cation.

Gating

The opening and closing of most channels is regulated by biological signals (Fig. 1) such as binding of intracellular or extracellular ligands (ligand-gated channels), changes in membrane potential (voltage-gated channels), changes in temperature²⁰ or mechanical stress²¹. In many cases, gating is also influenced by biochemical reactions such as phosphorylation, a process known as modulation.

Structural studies have revealed that most channels are closed by a gate that acts as a physical barrier to ion movement. This is achieved in a variety of ways. The gate of the gap-junction channel acts like the iris diaphragm of a camera¹⁵. In pore-loop channels, the inner TMs that line the pore act as a hinged gate, bending outwards to open the channel^{22,23}. Straightening of the helix and sliding of the intracellular ends together (at the helix bundle-crossing region) closes the pore. Inactivation of some K_v channels is produced by a tethered N-terminal blocker, which swings into the pore and physically plugs it^{14,24}. All these mechanisms involve large-scale movements of the protein backbone, but in CIC channels the conformational change is apparently quite subtle and the rotation of a single amino acid is sufficient to block ion permeation (see p. 484).

A non-steric gating mechanism has been proposed for AChR channels²⁵, although the resolution of the channel structure is not yet good enough to confirm the idea. Rather than physical occlusion, the gate may be an energetic one that results from a 'hydrophobic wall' to ion permeation. In AChR, closure is accompanied by rotation of TM2, which brings hydrophobic leucine side chains into the pore¹³. It is argued that unhydrated ions are actually small enough to pass through the pore but they are prevented from doing so because the energetic cost of dehydration next to a 'greasy wall' is too high²⁵. (Remember that ions normally move through AChR in the hydrated state.)

These two types of gate are physically distinct from the selectivity filter. However, there is accumulating evidence that in some channels the selectivity filter itself may act as a second gate, which governs the rapid closings (flickers) characteristic of many ion channels²⁶.

Voltage gating

Voltage-gated channels are opened or closed by changes in transmembrane voltage. A fundamental requirement for voltage sensing is the transfer of electrical charge through the membrane electric field in response to transmembrane voltage changes. This 'gating charge' can arise from charged residues on the protein, for example in the fourth transmembrane segment (S4) (see p. 463), or from charged activator-ligands binding to sites deep within the protein, as in the CIC channels (see p. 484). In the S4 family, where it has been most studied, voltage sensitivity arises from the movement of three to four positively charged residues in S4 within the voltage field (see p. 463). Although the detailed nature of the S4 movement is currently controversial, there is wide agreement on the identity of the gating charges.

Ligand gating

The basic principle of ligand gating is that the ligand must bind differently to the open and the closed conformations. If it binds more tightly to the open state, the ligand will be a channel activator, whereas if binding to the closed configuration is stronger, it will be an inhibitor. There are several other features common to all ligand-gated channels. First, the ligand-binding site is almost invariably located at the interface between two adjacent subunits or domains. As described in the accompanying reviews, this is true for the acetylcholine-binding site

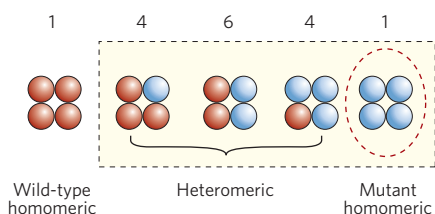
Box 1 | From malady to molecule and back again

New information on ion-channel structure affords unprecedented insight into how ion channels work and how disease-causing mutations achieve their functional effects. Conversely, identification of channel genes responsible for human disease has illuminated the relationship between channel structure and function. Kir6.2, the pore-forming subunit of the K_{ATP} channel (see p. 470), provides an example of this synergy.

The importance of the K_{ATP} channel in insulin secretion was established over 20 years ago⁵¹, and fuelled the search for K_{ATP} channel mutations causing diabetes. Recent studies show that heterozygous mutations in Kir6.2 cause neonatal diabetes by impairing the ability of ATP to inhibit the channel^{34,38} (see p. 470). This increases the K_{ATP} current, suppressing electrical activity and insulin secretion. Severe mutations that reduce ATP inhibition strongly also produce extra-pancreatic features such as epilepsy, developmental delay, muscle weakness and dysmorphic features (DEND syndrome)^{34,38}. Presumably these mutations increase K_{ATP} currents enough to reduce electrical activity in nerve and muscle.

The severity of the disease is dictated by both the tetrameric nature of the channel and the functional consequence of the mutation³⁴. Mutations that disrupt ATP binding usually cause only neonatal diabetes. This is because, although the channel has four ATP-binding sites (one per Kir6.2 subunit), binding of just one ATP shuts the pore²⁷. Because mutant and wild-type subunits associate randomly, in the heterozygous state only one-sixteenth of channels will have strongly reduced ATP sensitivity (dotted line in the figure). By contrast, the more severe DEND mutations usually stabilize the intrinsic open state of the channel and reduce ATP block indirectly³⁸ (see p. 470). Because each subunit can contribute to gating of the tetrameric pore, heteromeric channels containing both wild-type and mutant subunits will be affected (that is, as many as 15 out of 16 channels in heterozygotes; dashed line in the figure). Consequently DEND mutations produce a greater reduction in ATP sensitivity and more severe disease³⁸. Thus, Kir6.2 mutations illustrate how the tetrameric nature of the channel can result in very different phenotypes.

The different mechanisms of action of Kir6.2 mutations have important implications for therapy. In the past, neonatal diabetes was treated by insulin injection but many patients have now been successfully transferred to sulphonylurea tablets (which block K_{ATP} channels)³⁴ (see p. 470). However, these drugs (like ATP) are less effective on DEND mutations, which stabilize the channel open state. So the choice of therapy is dictated by the functional properties of the channel.



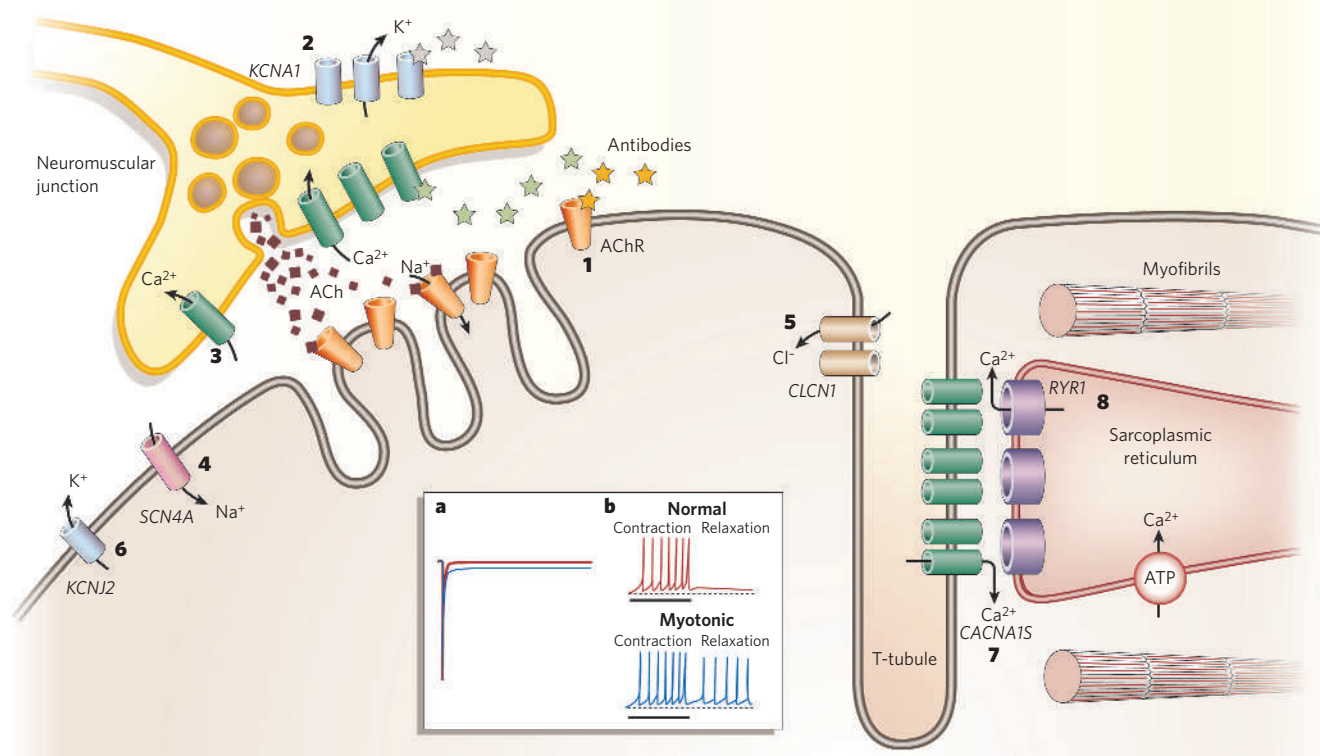


Figure 3 | Skeletal muscle channelopathies. **a**, Schematic of the motor nerve terminal, the neuromuscular junction and a skeletal muscle fibre, showing ion channels (numbered 1–8) whose mutation leads to disease (Tables 1 and 2). Mutations in several AChR subunits (1) cause myasthenia (muscle weakness). Autoimmune channelopathies are indicated by binding of antibodies (stars), which leads to internalization and downregulation of channel number. Loss of presynaptic K⁺-channel function (K_v1.1, KCNA1) (2) leads to increased transmitter release and enhanced muscle contraction, whereas downregulation of presynaptic Ca²⁺ channels (3) or AChR (1) function causes myasthenia, by preventing neurotransmitter release, or binding, respectively. Gain-of-function mutations in the muscle Na⁺ channel (4, SCN4A), and loss-of-function mutations in ClC channels (5), cause hyperexcitability and myotonia, while loss-of-function mutations in Kir1.1 (6) cause hyperexcitability and myotonia. Mutations in muscle Ca_v channels (7) or RYR channels (8) impair Ca²⁺ release from intracellular stores, producing malignant hyperthermia or paralysis (Tables 1, 2). **b**, Comparison of wild-type and mutant (red trace) Na⁺ current showing the persistent inward current produced by mutations associated with myotonia and/or periodic paralysis. **c**, Unlike normal muscle (above), action-potential firing in myotonic muscle (below) continues after stimulation has ceased, producing a sustained contraction.

of AChR¹³ (see p. 448), the glutamate-binding site of GluR (see p. 456) and the ATP-binding site of Kir6.2 (see p. 470; ref. 28), as well as the cyclic-nucleotide-binding site of CNG channels⁹. This arrangement is energetically more favourable because domain interfaces undergo larger conformational changes on ligand binding. And large conformational changes are needed because the ligand-binding sites lie at some considerable distance from the channel gate (about 50 Å in AChR¹³).

Second, many ligand-gated channels have more than one ligand-binding site, reflecting their multimeric structure. Functionally, this is important as it provides the potential for co-operative interactions. Positive co-operativity is required for the steep ligand dependence of gating that produces a sharp 'on-off' switch of channel activity, as in AChR channels (which have two acetylcholine-binding sites, see p. 448), CNG channels (which have four cyclic-nucleotide-binding sites⁹) and BK channels (which have eight Ca²⁺-binding sites⁸). In other cases, however, ligand-binding sites do not show co-operative interactions. The four inhibitory ATP-binding sites of the K_{ATP} channel act independently and one molecule of ATP is sufficient to close the channel²⁷, probably because the channel needs to respond to changes in ATP in a gently graded fashion (see p. 470). As explained in Box 1, the presence of multiple binding sites is important when analysing how heterozygous mutations affect the behaviour of multimeric channels.

A final common feature of ligand-gated channels is that flexible loops (termed gating loops) link the ligand-binding domains to the pore. This is discussed in detail for Cys-loop receptors (see p. 448). Similar loops

also link the ligand-binding domains of GluR channels (see p. 456) and the ATP- and Gβγ-binding domain of Kir channels^{28,29} (see also p. 470) to the pore.

Accessory proteins

Although ion channels are often viewed as consisting only of pore-forming subunits, this picture is too simplistic. Most channels possess tightly associated accessory subunits, and some form part of much larger macromolecular complexes. Accessory subunits specify the location and abundance of ion channels in the plasma membrane, modulate their biophysical properties and fine-tune their sensitivity to physiological ligands and pharmacological agents. Thus mixing and matching of accessory subunits with different pore-forming subunits contributes to the diversity of ion channels.

Many accessory proteins act like molecular chaperones, facilitating the trafficking of pore-forming subunits to the plasma membrane. In their absence, channels are reduced in number (Ca_v; ref. 30) or totally absent (K_{ATP}; ref. 31). Other ancillary proteins anchor the pore-forming subunit in the membrane or target it to specific locations: rapsyn, for example, clusters AChR at excitatory synapses (see p. 448). Accessory proteins can also be co-opted to serve other functions. The SUR subunit of the K_{ATP} channel has a key role in the metabolic regulation of the channel by hydrolysing ATP³² and stimulating channel opening (see p. 470); it also confers sensitivity to therapeutic drugs (see p. 470). The MinK and MiRP proteins modulate the single-channel conduct-

Table 1 | Pore-loop channelopathies

Family	Protein	Gene	G/L	Disease and symptoms
Kir	Kir1.1 (ROMK)	<i>KCNJ1</i>	L	Bartter's syndrome (renal salt loss)
	Kir2.1 (IRK)	<i>KCNJ2</i>	L	Andersen's syndrome
	Kir6.2, K_{ATP} channel α -subunit	<i>KCNJ11</i>	L G	Congenital hyperinsulinism Neonatal diabetes; DEND syndrome
	SUR1, β -cell K_{ATP} channel β -subunit	<i>SUR1</i>	L	Congenital hyperinsulinism
	SUR2, cardiac K_{ATP} channel β -subunit	<i>SUR2</i>	L	Dilated cardiomyopathy
K_V	$K_V1.1$, neuronal α -subunit	<i>KCNA1</i>	L	Episodic ataxia type 1, myokymia, neuromyotonia
	KVLQT1, neuronal α -subunit	<i>KCNQ1</i>	L G	Long QT syndrome 1 and possible deafness; Atrial fibrillation; Jervell-Lange-Neilsen syndrome Short QT syndrome
	KCNQ2, neuronal α -subunit	<i>KCNQ2</i>	L	Benign neonatal febrile convulsions
	KCNQ3, neuronal α -subunit	<i>KCNQ3</i>	L	Benign neonatal febrile convulsions
	KCNQ4, neuronal α -subunit	<i>KCNQ4</i>	L	Nonsyndromic deafness
	hERG, I_{Kr} cardiac α -subunit	<i>KCNH2</i>	L G	Long QT syndrome 2 Short QT syndrome
	MinK, I_{Ks} cardiac β -subunit	<i>KCNE1</i>	L	Long QT syndrome 5 Jervell-Lange-Neilsen syndrome
	MiRP1, I_{Kr} cardiac β -subunit	<i>KCNE2</i>	L	Long QT syndrome 6
	TRPP2 (Polycystin 2, PKD2)	<i>TRPP2</i>		Autosomal dominant polycystic kidney disease
	TRPC6	<i>TRPC6</i>	G	Focal segmental glomerulosclerosis, defective Mg reabsorption
	TRPML1 (Mucolipin)	<i>Mcoln1</i>		Mucopolidosis IV
CNG	Retinal cGMP gated $\alpha 1$ -subunit	<i>CNGA1</i>	L	Retinitis pigmentosa
	Retinal cGMP gated $\alpha 3$ -subunit	<i>CNGA3</i>	L	Achromatopsia-2
	Retinal cGMP gated $\beta 3$ -subunit	<i>CNGB3</i>	L	Achromatopsia-3
K_{Ca}	BK channel α -subunit	<i>KCMNA1</i>	G	Generalized epilepsy with paroxysmal dyskinesia
Na_V	$Na_V1.1$, neuronal α -subunit	<i>SCN1A</i>	G L	Generalized epilepsy with febrile seizures type 2 Severe myoclonic epilepsy of infancy
	Neuronal β -subunit	<i>SCN1B</i>	L	Generalized epilepsy with febrile seizures type 1
	$Na_V2.1$, neuronal α -subunit	<i>SCN2A</i>	G	Benign familial neonatal seizures
	Skeletal muscle α -subunit	<i>SCN4A</i>	G L	Paramyotonia congenita; Hyperkalemic periodic paralysis; K^+ -aggravated myotonia Hypokalemic periodic paralysis
	Cardiac muscle α -subunit	<i>SCN5A</i>	G	Long QT syndrome 2; Brugada syndrome; Non-progressive congenital heart block; Progressive cardiac conduction defect
		<i>SCN9A</i>	G	Familial erythralgia (pain)
Ca_V	Neuronal α -subunit	<i>CACNA1A</i>	L G E	Episodic ataxia type 2 Familial hemiplegic migraine Spinocerebellar ataxia type 6
	Retinal α -subunit	<i>CACNA1F</i>		Congenital stationary night blindness
	Skeletal muscle α -subunit	<i>CACNA1S</i>		Hypokalemic periodic paralysis; malignant hypothermia type 5
	Neuronal $\beta 4$ -subunit	<i>CACNB4</i>		Juvenile myoclonic epilepsy; Generalized epilepsy and praxis-induced seizures; Episodic ataxia type 3
	$Ca_V1.2$, α -subunit	<i>CACNA1C</i>		Timothy syndrome

G, gain of function; L, loss of function; E, repeat expansion. The symptom of long and short QT syndrome is ventricular arrhythmia. The symptoms of Jervell-Lange-Neilsen syndrome include deafness and cardiac arrhythmias. Symptoms of autosomal dominant polycystic kidney disease include cardiac septal defects. Achromatopsia is total colour blindness. The symptoms of Timothy syndrome are multi-organ dysfunction including long QT syndrome and syndactyly. DEND syndrome is characterized by developmental delay, epilepsy, muscle weakness and neonatal diabetes. Anderson's syndrome is a multi-organ disorder that includes potassium-sensitive periodic paralysis and ventricular arrhythmia. Brugada syndrome results in ventricular fibrillation. For references, see: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM&cmd=Limits> (type in the gene name).

ance, gating and pharmacology of certain kinds of K_V channel (see p. 463), and the inactivation properties of many other voltage-gated channels originate in their accessory subunits^{2,30}. Accessory subunits may even serve as kinases that phosphorylate the pore-forming subunit or other proteins³³.

For these reasons, mutations in accessory subunits produce a variety of channelopathies (Tables 1 and 2). For example, mutations in rapsyn decrease AChR density, producing congenital myasthenic syndrome (see p. 448); loss-of-function mutations in SUR cause congenital hyperinsulinaemia by decreasing K_{ATP} channel activity (see p. 470 and ref. 34); and mutations in MinK and MiRP lead to long QT (LQT) syndrome, which can cause fatal arrhythmia (see p. 463) (Table 1). Ancillary subunits are also the targets of many therapeutic drugs (see p. 470), making them of key importance to the pharmaceutical industry.

Concepts in channelopathies

The myriad diseases arising from impairment of channel function exemplify the importance of ion channels to the organism (Tables 1 and 2). The range of channelopathies is too great, and too diverse, to cover in a short review — it needs a book¹. Therefore I shall focus on just a few key concepts or examples that are helpful for understanding the reviews in this Insight. Further information can be found at the Online Mendelian Inheritance in Man website (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM&cmd=Limits).

Disorders caused by defective ion-channel function result from several different causes but the term channelopathies is usually confined to disorders resulting from mutations in ion-channel genes themselves (Tables 1 and 2). To date, mutations in over 60 ion-channel genes have been associated with human disease. Inherited channelopathies are usu-

Table 2 | Non-pore-loop channelopathies

Channel protein	Gene	G/L	Disease
Epithelial Na ⁺ channel (ENaC) β-subunit	<i>SCNN1B</i>	G L	Hypertension (Liddle's syndrome) Hypotension (pseudohypoaldosteronism type 1)
Epithelial Na ⁺ channel (ENaC) γ-subunit	<i>SCNN1G</i>	G L	Hypertension (Liddle's syndrome) Hypotension (pseudohypoaldosteronism type 1)
Skeletal muscle sarcoplasmic reticulum (SR) calcium release channel	<i>RYR1</i>	G G	Malignant hyperthermia (triggered by anaesthetics and muscle relaxants) Central core disease (muscle weakness)
Cardiac SR calcium release channel	<i>RYR2</i>	G	Cardiac arrhythmia (catecholaminergic polymorphic VT)
CFTR, epithelial Cl ⁻ channel	<i>CFTR</i>	L	Cystic fibrosis
Bestrophin, epithelial Cl ⁻ channel β-subunit	<i>BEST1</i>	L	Vitelliform macular dystrophy (Best disease)
CIC1, Cl ⁻ channel skeletal muscle α-subunit	<i>CLCN1</i>	L L	Generalized myotonia (Becker's disease) Myotonia congenita (Thomson's disease)
CLC2, Cl ⁻ channel α-subunit	<i>CLCN2</i>	L	Several types of epilepsy
CIC5, Cl ⁻ transporter kidney	<i>CLCN5</i>	L	Dent's disease (X-linked) and other renal tubular disorders
CIC7, Cl ⁻ transporter	<i>CLCN7</i>	L	Osteopetrosis (dense bones), sometimes with blindness
CIC-Ka, Cl ⁻ channel kidney α-subunit	<i>CLCNKA</i>	L	Bartter's syndrome
CIC-Kb, Cl ⁻ channel kidney α-subunit	<i>CLCNKB</i>	L	Bartter's syndrome
Barttin, Cl ⁻ channel β-subunit		L	Bartter's syndrome with deafness
Cys-loop receptors			
Glycine receptor neuronal α-subunit	<i>GLRA1</i>	L	Hyperekplexia (stiff baby syndrome)
GABA _A receptor neuronal α1-subunit	<i>GABRA1</i>	L	Juvenile myoclonic epilepsy
AChR skeletal muscle α1-subunit	<i>CHRNA1</i>	G	Congenital myasthenia
AChR neuronal α4-subunit	<i>CHRNA4</i>	L	Autosomal dominant nocturnal frontal lobe epilepsy
AChR skeletal muscle β1-subunit	<i>CHRNA1</i>	G	Congenital myasthenia
AChR skeletal muscle δ-subunit	<i>CHRNA1</i>	G	Congenital myasthenia
AChR skeletal muscle ε-subunit	<i>CHRNA1</i>	G L	Congenital myasthenia Fast-channel syndrome (muscle weakness)
AChR neuronal β-subunit	<i>CHRNA1</i>	L	Autosomal dominant nocturnal frontal lobe epilepsy

G, gain of function. L, loss of function. Generalized myotonia (Becker's disease) and myotonia congenita (Thomson's disease) both result in skeletal muscle hyperexcitability. Bartter's syndrome results in renal salt loss. Congenital myasthenia results in muscle weakness. Hyperekplexia results in muscle contractions when startled. Mutations in CIC5 can cause kidney stones, hypophosphatemic rickets and low molecular weight proteinuria. For references see: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM&cmd=Limits>.

ally very rare. An exception is cystic fibrosis, which affects around 1 in 2,000 people in Northern Europe and the United States. This is a recessive disease, and as many as 1 in 20 people are carriers (see p. 477).

Many channelopathies are genetically heterogeneous, the same clinical phenotype being caused by mutations in different genes. For example, mutations in at least eight different genes give rise to LQT syndrome (Table 1). Furthermore, the same mutation may have greater or lesser effects on different genetic backgrounds. This is one reason why not all carriers of the same disease-causing mutation are affected or why they exhibit different phenotypes. About half the patients carrying the Thr857Met mutation in the Na_v channel α-subunit experience only childhood epilepsy, whereas the other half also have epilepsy as adults³⁵. Environmental effects may also influence the extent to which a phenotype is expressed.

Channel-mediated ion flux is determined by the number of channels in the membrane, the fraction of time they remain open (the open probability) and the conductance of the single channel. In principle, alteration of any of these could disrupt channel function, but the last is found only rarely (some CFTR mutations alter single-channel conductance³⁶). Loss-of-channel function is commonly the result of mutations that prevent protein synthesis or correct membrane targeting so that the channel density is lower, or that impair channel activity, by disrupting ligand binding, for example. Gain-of-function mutations may also produce disease. Indeed, gain- and loss-of-function mutations in the same gene can produce distinct clinical disorders: neonatal diabetes (due to too little insulin secretion) and congenital hyperinsulinaemia both result from mutations in Kir6.2 (see p. 470; ref. 34).

The form and severity of the clinical phenotype manifested by individuals carrying a given ion-channel mutation depends on several factors. The extent to which the mutation modifies channel function is

obviously crucial. However, the relative contribution the channel makes to the electrical activity of the cell is also important. Because cells in different tissues have variable complements of ion channels, their electrical activity can be affected differently. For example, K_{ATP} channels are found in tissues as diverse as brain, muscle and pancreatic beta-cells, but only the most severe mutations produce extra-pancreatic effects³⁷, because only in beta-cells is the resting potential determined predominantly by the K_{ATP} channel³⁸.

Different mutations in the same gene may give rise to either recessive or dominant disease, as is the case for CIC1 mutations that cause both myotonia congenita (dominant) and generalized myotonia (recessive)³⁹. A dominant disease results if more than 50% of the channel protein is required for normal function (haploinsufficiency) or if the mutant protein interferes with the functional activity of the wild-type protein (the dominant-negative effect), as occurs when a single mutant subunit in a multimeric channel prevents its function. Because heterozygous individuals will express both wild-type and mutant channels in the same cell, heteropolymerization can result in non-functional channels, causing dominant disease. However, a small percentage of homomeric wild-type channels will be present, which may explain why dominant mutations are rarely as severe as recessive ones. Whether or not a mutation is dominant or recessive thus depends on whether the channel is multimeric and if the presence of one (or more) mutant subunit affects the wild-type subunits in the complex. The multimeric nature of ion channels may also influence the severity of recessive disorders (Box 1).

The perils of over- or under-excitement

Ion channels are essential for membrane electrical excitability, and their mutation results in many nerve, muscle and endocrine disorders (Fig. 3). Na_v channels mediate the upstroke of the action potential in

nerve and muscle, first opening in response to depolarization and then rapidly entering a non-conducting inactivated state^{2,14} (Fig. 1b). Most mutations in Na_v-channel genes act as gain-of-function mutations³⁵. They slow or impair inactivation, which leads to a persistent Na⁺ current and membrane hyperexcitability (Fig. 3a, b). In central neurons, such hyperexcitability results in epilepsy³⁵; in skeletal muscle it causes myotonia³⁹ (prolonged muscle contraction); and in the heart it causes LQT syndrome (see p. 463) (Tables 1). It is worth noting that the change in channel activity required to produce disease can be very subtle. A persistent Na⁺ current just 2–5% of the maximum is sufficient to cause epilepsy or muscle disorders⁴¹.

Action-potential repolarization is mediated by K_v channels. Unsurprisingly, therefore, loss-of-function mutations in K_v channel subunits prolong the action-potential duration, inducing a hyperexcitability that leads to epilepsy, episodic ataxia and myokymia as well as cardiac disorders (see p. 463; ref. 40; Table 1). Paradoxically, gain-of-function mutations in K⁺ channels can also cause cardiac arrhythmia (short QT syndrome; see p. 463) or epilepsy^{34,38,41}. The latter may result if the mutation preferentially reduces the excitability of inhibitory neurons³⁴. Faster action-potential repolarization may also enable Na_v channels to recover more quickly from inactivation, thereby increasing the firing rate⁴².

Epithelial channelopathies

Ion channels are important for ion fluxes across epithelial membranes and thus for salt and water balance. Mutations in the epithelial Na⁺ channel (ENaC) lead to enhanced, or reduced, Na⁺ uptake in the kidney tubules⁴². Because plasma Na⁺ levels influence blood pressure, this causes hereditary hypertension or hypotension, respectively. Mutations in epithelial Cl[−] channels⁴³, or Kir channels⁴⁴, also cause renal disorders (Table 2), and cystic fibrosis results from impaired fluid secretion, particularly in the airways (see p. 477). The importance of certain K_v channels in fluid secretion in the cochlear was first recognized from the fact that their mutation causes deafness⁴⁰.

Accumulating evidence suggests that ion channels have a role in cell proliferation and cancer. For example, K_v1.3 triggers lymphocyte activation and is overexpressed in various tumours, and ectopic expression of the K_v channel Eag1 outside the nervous system triggers tumour formation⁴⁵.

Intracellular ion channels

Ion channels are not confined to the plasma membrane. They are found in the membranes of all intracellular organelles, including secretory and endocytotic vesicles, synaptic vesicles, lysosomes, mitochondria and the endoplasmic and sarcoplasmic reticulum (SR). Not surprisingly, mutations in their genes also cause disease. Thus, mutations in the SR ryanodine receptor, which regulates Ca²⁺ release from intracellular stores, lead to malignant hyperthermia, a disorder in which common inhalation anaesthetics produce uncontrolled muscle contractions and a potentially fatal rise in body temperature⁴⁶. The eukaryotic cousins of the Cl[−] transporters described by Miller in this issue (p. 484) are crucial for acidification of endocytotic vesicles in epithelial tissues⁴³. Their mutation can produce osteopetrosis, in which the bone marrow is replaced by bone, or Dent's disease, which is characterized by low-molecular-weight proteinuria, kidney stones, hypercalciuria, nephrocalcinosis and occasionally renal failure⁴³.

Other types of channel disease

Although not strictly considered as channelopathies, mutations in other genes can adversely affect ion-channel function. These include genes involved in trafficking, localizing and anchoring channels in the plasma membrane or those that influence the concentration of ligands or modulators of channel function. Congenital myasthenia, for example, can be produced by mutations in rapsyn, by a deficiency of acetylcholinesterase or acetylcholine production, as well as by mutations in the AChR itself (see p. 448). Similarly, mutations in regulatory genes can cause diabetes³⁷ by affecting the activity of ion channels in pancreatic beta-cells. Autoantibodies against ion channels also cause a range of

disorders, including myasthenia gravis⁴⁷, Lambert–Eaton myasthenic syndrome⁴⁷, acquired neuromyotonia^{47,48} and CNS disorders⁴⁸. These are sometimes known as the autoimmune channelopathies. Ion channels are also the target of a large and diverse group of toxins that bind to ion channels and either inhibit or enhance their activity¹⁴. In addition, therapeutic drugs can cause iatrogenic disease if they interact adversely with ion channels. These are sometimes referred to as acquired channelopathies. The best-known example is that of drugs that inhibit the cardiac K_v channel hERG and thereby produce acquired LQT syndrome (see p. 463). Interestingly, specific polymorphisms (gene variants) seem to predispose certain individuals to drug-induced LQT⁴⁹.

Future perspectives

Almost 350 years ago, William Harvey wrote: “Nor is there any better way to advance the proper practice of medicine than ... by careful investigation of the rarer forms of disease”. Studies of channelopathies illustrate this maxim's current relevance, for they have identified novel ion-channel genes, illuminated relationships between channel structure and function, and revealed diverse and unexpected physiological roles for ion channels. And, importantly, they have also led to new therapeutic practice, to new diagnostic tests for genetic disease, to routine screening of drugs to prevent acquired channel disorders and to novel drug therapy.

Ion-channel research is now poised at an exciting point. In the next decade, we anticipate new high-resolution structures and rapid advances in our understanding of how ion channels work. With luck, we will discover the key structural elements that distinguish ion channels from their cousins the transporters. A better understanding of how gating works, how accessory subunits regulate channel structure and how lipids interact with ion channels is likely. Advances are also expected from studies of channelopathies, both in terms of structure–function relationships and of how channels contribute to cellular and systems physiology. A particular challenge will be to elucidate the extent to which polymorphisms in ion-channel genes predispose individuals towards common polygenic diseases such as type 2 diabetes, hypertension, obesity and cardiac disease. At least in the case of diabetes³⁷ and hypertension³⁰ there is evidence that this is the case. We can also expect that advances in ion-channel research will lead to new therapies, in the form of new drugs and of drugs tailored to the individual's genetic profile. ■

1. Ashcroft, F. M. *Ion Channels and Disease* (Academic Press, New York, 2000).
2. Ulbricht, W. Sodium channel inactivation: molecular determinants and modulation. *Physiol. Rev.* **85**, 1271–1301 (2005).
3. Johnston, G. A. GABA(A) receptor channel pharmacology. *Curr. Pharm. Des.* **11**, 1867–1868 (2005).
4. Fischer, W. B. & Sansom, M. S. Viral ion channels: structure and function. *Biochim. Biophys. Acta* **1561**, 27–45 (2002).
5. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* **280**, 69–77 (1998).
6. Kuo, A. *et al.* Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science* **300**, 1922–1926 (2003).
7. Long, S. B., Campbell, E. B. & Mackinnon, R. Crystal structure of a mammalian voltage-dependent Shaker family K⁺ channel. *Science* **309**, 897–903 (2005).
8. Magleby, K. L. Gating mechanism of BK (Slo1) channels: so near, yet so far. *J. Gen. Physiol.* **121**, 81–96 (2003).
9. Matulef, K. & Zagotta, W. N. Cyclic nucleotide-gated ion channels. *Annu. Rev. Cell Dev. Biol.* **19**, 23–44 (2003).
10. Brejc, K. *et al.* Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. *Nature* **411**, 269–276 (2001).
11. Tasneem, A., Iyer, L. M., Jakobsson, E. & Aravind, L. Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels. *Genome Biol.* **6**, R4 (2005).
12. Dutzler, R., Campbell, E. B. & MacKinnon, R. Gating the selectivity filter in ClC chloride channels. *Science* **300**, 108–112 (2003).
13. Unwin, N. Refined structure of the nicotinic acetylcholine receptor at 4 Å resolution. *J. Mol. Biol.* **346**, 967–989 (2005).
14. Hille, B. *Ionic Channels of Excitable Membranes* 3rd edn (Sinauer, Massachusetts, 2001).
15. Unwin, P. N. & Ennis, P. D. Two configurations of a channel-forming membrane protein. *Nature* **307**, 609–613 (1984).
16. Wei, C. J., Xu, X. & Lo, C. W. Connexins and cell signaling in development and disease. *Annu. Rev. Cell Dev. Biol.* **20**, 811–838 (2004).
17. Heginbotham, L., Lu, Z., Abramson, T. & MacKinnon, R. Mutations in the K⁺ channel signature sequence. *Biophys. J.* **66**, 1061–1067 (1994).
18. Yang, J., Ellinor, P. T., Sather, W. A., Zhang, J. F. & Tsien, R. W. Molecular determinants of Ca²⁺ selectivity and ion permeation in L-type Ca²⁺ channels. *Nature* **366**, 158–161 (1993).
19. Heinemann, S. H., Terlau, H., Stühmer, W., Imoto, K. & Numa, S. Calcium channel characteristics conferred on the sodium channel by single mutations. *Nature* **356**, 441–443 (1992).

20. Ramsey, I. S., Delling, M. & Clapham, D. E. An Introduction to TRP Channels. *Annu. Rev. Physiol.* **68**, 619–647 (2006).
21. Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. & Rees, D. C. Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel. *Science* **282**, 2220–2226 (1998).
22. Jiang, Y. *et al.* Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**, 515–522 (2002).
23. Webster, S. M., Del Camino, D., Dekker, J. P. & Yellen, G. Intracellular gate opening in Shaker K⁺ channels defined by high-affinity metal bridges. *Nature* **428**, 864–868 (2004).
24. Kukuljan, M., Labarca, P. & Latorre, R. Molecular determinants of ion conduction and inactivation in K⁺ channels. *Am. J. Physiol.* **268**, C535–C556 (1995).
25. Amiri, S., Tai, K., Beckstein, O., Biggin, P. C. & Sansom, M. S. P. The $\alpha 7$ nicotinic acetylcholine receptor: molecular modelling electrostatics, and energetics. *Mol. Membr. Biol.* **22**, 151–162 (2005).
26. Proks, P., Capener, C., Jones, P. & Ashcroft, F. M. Mutations within the P-loop of Kir6.2 modulate the intraburst kinetics of the ATP-sensitive potassium channel. *J. Gen. Physiol.* **118**, 341–353 (2001).
27. Markworth, E., Schwanstecher, C. & Schwanstecher, M. ATP4⁻ mediates closure of pancreatic beta-cell ATP-sensitive potassium channels by interaction with 1 of 4 identical sites. *Diabetes* **49**, 1413–1418 (2000).
28. Antcliff, J. F., Haider, S., Proks, P., Sansom, M. S. & Ashcroft, F. M. Functional analysis of a structural model of the ATP-binding site of the K_{ATP} channel Kir6.2 subunit. *EMBO J.* **24**, 229–239 (2005).
29. Pegan, S. *et al.* Cytoplasmic domain structures of Kir2.1 and Kir3.1 show sites for modulating gating and rectification. *Nature Neurosci.* **8**, 279–287 (2005).
30. Richards, M. W., Butcher, A. J. & Dolphin, A. C. Ca²⁺ channel beta-subunits: structural insights AID our understanding. *Trends Pharmacol. Sci.* **25**, 626–632 (2004).
31. Zerangue, N., Schwappach, B., Jan, Y. N. & Jan, L. Y. A new ER trafficking signal regulates the subunit stoichiometry of plasma membrane K(ATP) channels. *Neuron* **22**, 537–548 (1999).
32. Mikhailov, M. V. *et al.* 3D structural and functional characterization of the purified K_{ATP} channel complex Kir6.2-SUR1. *EMBO J.* **24**, 4166–4175 (2005).
33. Cai, S. Q., Hernandez, L., Wang, Y., Park, K. H. & Sesti, F. MPS-1 is a K⁺ channel beta-subunit and a serine/threonine kinase. *Nature Neurosci.* **8**, 1503–1509 (2005).
34. Ashcroft, F. M. ATP-sensitive Potassium Channelopathies: focus on insulin secretion. *J. Clin. Invest.* **115**, 2047–2058 (2005).
35. Meisler, M. H. & Kearney, J. A. Sodium channel mutations in epilepsy and other neurological disorders. *J. Clin. Invest.* **115**, 2010–2017 (2005).
36. Sheppard, D. N. *et al.* Mutations in CFTR associated with mild-disease-form Cl⁻ channels with altered pore properties. *Nature* **362**, 160–164 (1993).
37. Proks, P. *et al.* Molecular basis of Kir6.2 mutations associated with neonatal diabetes or neonatal diabetes plus neurological features. *Proc. Natl Acad. Sci. USA* **101**, 17539–17544 (2004).
38. Ashcroft, F. M. & Rorsman, P. Type-2 diabetes mellitus: not quite exciting enough? *Hum. Mol. Genet.* **13**, R21–R31 (2004).
39. Jurkat-Rott, K. & Lehmann-Horn, F. Muscle channelopathies and critical points in functional and genetic studies. *J. Clin. Invest.* **115**, 2000–2009 (2005).
40. Jentsch, T. J. Neuronal KCNQ potassium channels: physiology and role in disease. *Nature Rev. Neurosci.* **1**, 21–30 (2000).
41. Du, W. *et al.* Calcium-sensitive potassium channelopathy in human epilepsy and paroxysmal movement disorder. *Nature Genet.* **37**, 733–738 (2005).
42. Rossier, B. C., Pradervand, S., Schild, L. & Hummler, E. Epithelial sodium channel and the control of sodium balance: interaction between genetic and environmental factors. *Annu. Rev. Physiol.* **64**, 877–897 (2002).
43. Jentsch, T. J., Maritzen, T. & Zdebik, A. A. Chloride channel diseases resulting from impaired transepithelial transport or vesicular function. *J. Clin. Invest.* **115**, 2039–2046 (2005).
44. Simon, D. B. *et al.* Genetic heterogeneity of Bartter's syndrome revealed by mutations in the K⁺ channel, ROMK. *Nature Genet.* **14**, 152–156 (1996).
45. Pardo, L. A., Contreras-Jurado, C., Zientkowska, M., Alves, F. & Stuhmer, W. Role of voltage-gated potassium channels in cancer. *J. Membr. Biol.* **205**, 115–124 (2005).
46. Priori, S. G. & Napolitano, C. Cardiac and skeletal muscle disorders caused by mutations in the intracellular Ca²⁺ release channels. *J. Clin. Invest.* **115**, 2033–2038 (2005).
47. Lang, B. & Vincent, A. Autoantibodies to ion channels at the neuromuscular junction. *Autoimmun. Rev.* **2**, 94–100 (2003).
48. Newsom-Davis, J. *et al.* Autoimmune disorders of neuronal potassium channels. *Ann. NY Acad. Sci.* **998**, 202–210 (2003).
49. Sesti, F. *et al.* A common polymorphism associated with antibiotic-induced cardiac arrhythmia. *Proc. Natl Acad. Sci. USA* **97**, 10613–10618 (2000).
50. Ambrosius, W. T. *et al.* Genetic variants in the epithelial sodium channel in relation to aldosterone and potassium excretion and risk for hypertension. *Hypertension* **34**, 631–637 (1999).
51. Ashcroft, F. M., Harrison, D. E. & Ashcroft, S. J. H. Glucose induces closure of single potassium channels in isolated rat pancreatic β -cells. *Nature* **312**, 446–448 (1984).

Acknowledgements I thank C. Miller and M. Sansom for comments and criticisms on earlier versions of this review. Work in my laboratory is supported by grants from the Royal Society, the Wellcome Trust and the European Union (BioSim and EuroDia).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to F.A. (frances.ashcroft@physiol.ox.ac.uk).

Recent advances in Cys-loop receptor structure and function

Steven M. Sine¹ & Andrew G. Engel²

Throughout the nervous system, moment-to-moment communication relies on postsynaptic receptors to detect neurotransmitters and change the membrane potential. For the Cys-loop superfamily of receptors, recent structural data have catalysed a leap in our understanding of the three steps of chemical-to-electrical transduction: neurotransmitter binding, communication between the binding site and the barrier to ions, and opening and closing of the barrier. The emerging insights might be expected to explain how mutations of receptors cause neurological disease, but the opposite is generally true. Namely, analyses of disease-causing mutations have clarified receptor structure–function relationships as well as mechanisms governing the postsynaptic response.

When an action potential (AP) depolarizes a nerve terminal, neurotransmitter released from storage vesicles diffuses a short distance to the postsynaptic membrane, where it binds to its receptor and opens an intrinsic channel through which small ions flow. If the receptor channel is cation-selective the cell depolarizes and a new AP is generated, whereas if it is anion selective the cell hyperpolarizes and a larger stimulus is required to elicit the AP. The process from neurotransmitter release to the postsynaptic response occurs in less than a millisecond, making receptor-coupled channels ideal for moment-to-moment communication.

Receptors of the Cys-loop superfamily constitute a major class of receptor-coupled ion channels. They contain five protein subunits, each harbouring a signature sequence of 13 residues flanked by cysteines, which bond covalently to form a closed loop situated between binding and channel domains. Cys-loop receptors are further subdivided into cation- and anion-selective channels, corresponding to excitatory receptors activated by acetylcholine (ACh) or serotonin (5-hydroxytryptamine or 5-HT), and inhibitory receptors activated by GABA (γ -aminobutyric acid) or glycine. Both cationic and anionic receptors can assemble from five copies of a single type of subunit, but they more commonly assemble from several different types of subunit. Formation of such hetero-pentamers, together with the many subunit types, provides a wealth of functional diversity, presumably to meet a wide range of synaptic needs.

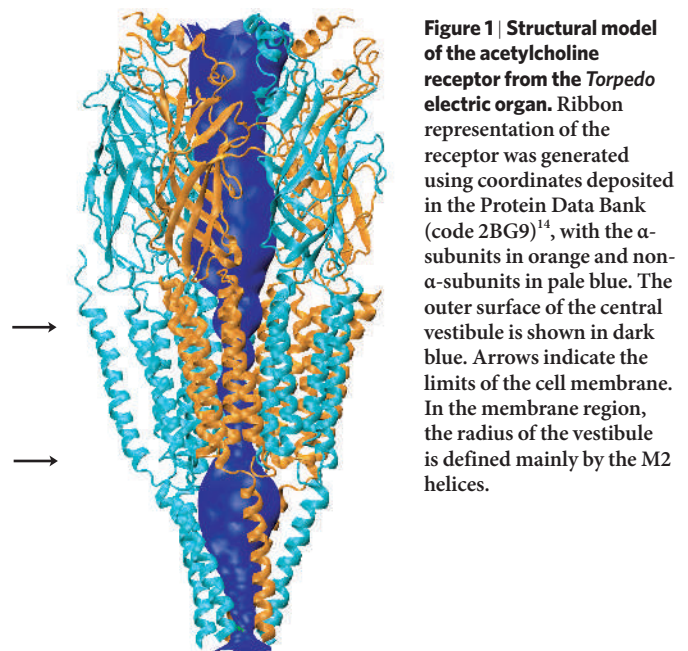
Here we describe recent advances in our understanding of the chemical-to-electrical transduction mediated by Cys-loop receptors. Transduction is divided into three processes: neurotransmitter binding, channel gating, and coupling of binding to gating. Our discussion focuses on nicotinic acetylcholine receptors (AChRs) from skeletal muscle that mediate voluntary movement, and on how analyses of disease-causing mutations clarify both the transduction process and mechanisms governing the postsynaptic response.

Overview of structure and mechanism

Ever since the application of ACh was shown to rapidly increase membrane conductance at the motor endplate¹, synaptic receptors have been viewed as a binding site and ion channel within the same macromolecule.

Purification and reconstitution of receptors from the *Torpedo* electric ray established the continuity of binding site and channel domains², and electron microscopy of two-dimensional arrays of the receptor revealed its cylindrical shape and dimensions, defined subunit boundaries, and provided glimpses of the internal structures of the binding sites and channel^{3,4}. This supra-molecular view of the receptor required several decades of experimental work, but by the new millennium a plateau had been reached in understanding receptor structure.

However, two breakthroughs propelled the field into the atomic structural age. The first was the discovery of acetylcholine-binding protein (AChBP) from the freshwater snail *Lymnaea stagnalis*⁵, and the determination of its structure at a resolution of 2.7 Å (ref. 6). AChBP is a

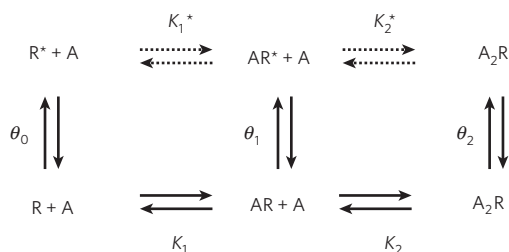


pentamer of identical subunits, each homologous to the amino-terminal domain of a Cys-loop receptor subunit, and contains structural cornerstones that give Cys-loop receptors their unique signature. These include a Cys-loop, a size and shape mimicking the synaptic protrusion of the *Torpedo* receptor, ACh-binding sites formed at interfaces between subunits, and a cluster of aromatic residues ideal for stabilizing the cationic ACh. Opportunities for advancing Cys-loop receptor structure arose immediately, such as homology modelling of ligand-binding domains using AChBP as the template^{7–10}. Moreover, studies of structure–function relationships no longer relied on one-dimensional sequence alignments to generate hypotheses. Instead, pointed hypotheses originated from three-dimensional structural information.

The second breakthrough came in two stages, both from electron microscopy of two-dimensional arrays of *Torpedo* receptors at a resolution of 4 Å, which revealed the protein backbone, and locations of α -carbon atoms and bulky side chains, but resolution of small side chains and side-chain orientations was limited. The first stage was an atomic-scale model of the receptor channel, which revealed four transmembrane α -helices (M1–M4) per subunit and their positioning within the pentameric channel¹¹. A second round of homology modelling ensued using AChBP and the *Torpedo* channel as templates^{12,13}, allowing the coupling mechanism between binding site and channel to be investigated in unprecedented detail. The second stage was an atomic-scale model of the *Torpedo* receptor encompassing the majority of the binding domain, the channel domain and about half of the cytoplasmic domain¹⁴. Although the resolution was lower than that for AChBP, the overall structural model revealed the spatial relationship between the binding and channel domains, which differed from that in homology models derived from separate AChBP and channel templates. Compared with homology models, the experimentally determined structural model showed greater separation between binding and channel domains, and the rotation of one domain relative to the other differed, allowing new and more precise hypotheses about the atomic-scale connections that functionally couple the two domains.

The emerging structure of a Cys-loop receptor comprises an extracellular domain of mainly β -sheets, a channel domain of α -helices, and a cytoplasmic domain containing substantial α -helix (Fig. 1). Each subunit spans the full length of the receptor and is positioned approximately normal to the membrane, contributing to all three structural domains. This subunit arrangement gives rise to a vestibule running through the receptor, which allows flow of ions and gating of the flow by dilation of its narrowest constriction.

Binding of neurotransmitter triggers a step increase in current through the receptor channel, which persists while neurotransmitter is bound but terminates when the channel closes and neurotransmitter dissociates. Summation of many such unitary current pulses generates the postsynaptic response (Fig. 2). How the neurotransmitter opens the channel has been the subject of several decades of research, with the greatest advances emerging from mechanistic studies of AChRs from skeletal muscle^{15–18}. In the absence of agonist, the AChR channel can open but with a vanishingly low probability¹⁹. So, activation by ACh can be viewed as an amplification process in which the probability that the channel will open is greatly increased. For all Cys-loop receptors, at least two bound agonists are required for full amplification, leading to the following mechanism:



Here the receptor in the closed state (R) switches to the open state (R^*) in a reversible process called gating. Two agonist molecules (A) bind to the closed state, and the gating equilibrium constant θ increases as the

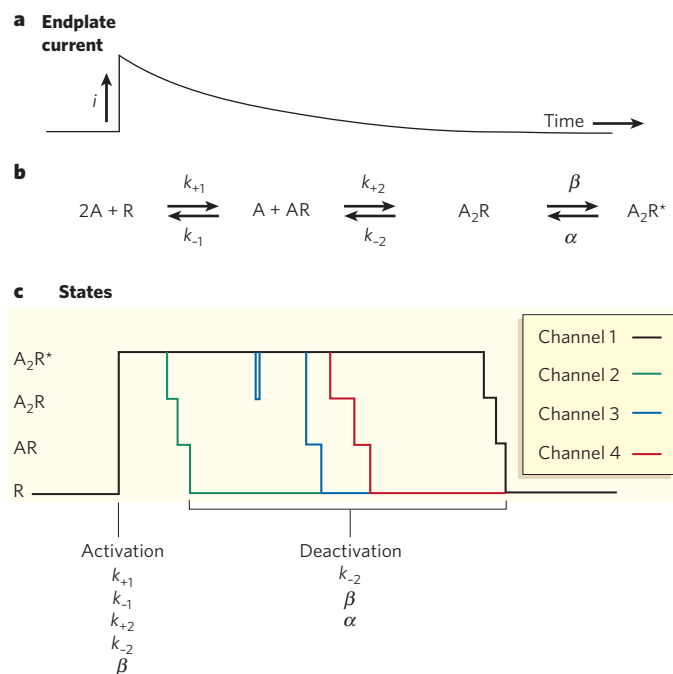


Figure 2 | Mechanism of activation of acetylcholine receptors at the motor endplate. **a**, Time course of an idealized endplate current, resulting from synchronous activation of many receptors by a transient pulse of nerve-released acetylcholine (ACh), showing the instantaneous rise and exponential decay. **b**, Kinetic description of receptor activation with microscopic rate constants for each step labelled. For clarity, a subset of states from the text scheme is shown, where R is the receptor in the closed state, R^* is the receptor in the open state, A is ACh, k_+ and k_- are ACh association and dissociation rate constants, respectively, and β and α are the channel opening and closing rate constants, respectively. **c**, Hypothetical state transitions of four receptor channels, each a different colour, that activate synchronously to produce the endplate current. Rate constants that govern the rise (activation) and decay (deactivation) of the endplate current are indicated below. Slow-channel congenital myasthenic syndrome (CMS) mutations affect mainly rate constants governing deactivation, whereas fast-channel CMS mutations can affect activation, deactivation or both.

number of bound agonist molecules increases. The amplification of θ by agonist is truly astounding — estimated to be some ten-million-fold with two agonists. The origin of this amplification lies in a much greater affinity of agonist for the open than for the closed state. For ACh, dissociation constants for the closed state, K_1 and K_2 , are in the micromolar range, whereas those for the open state, K_1^* and K_2^* , although not directly measurable are estimated to be in the nanomolar range^{20,21}. So, mechanistic considerations predict that the conformation of the binding site should change markedly between closed and open channel states.

In voluntary skeletal muscle, ACh associates rapidly, around tenfold slower than diffusion, and if the receptor channel does not open, ACh dissociates in a fraction of a millisecond^{22–24}. However, if the channel opens while ACh is bound, the agonist remains bound owing to high affinity of the open state. Once the channel closes, reopening can occur without dissociation of ACh, but this depends on the rate of channel opening relative to that for dissociation²⁵; for the muscle AChR, although ACh dissociates rapidly, channel opening is roughly twice as rapid, giving several re-openings per ACh occupancy. So, the kinetics of ACh occupancy shape the endplate current (EPC) in two ways: rapid association contributes to the rapid rise, whereas rapid dissociation coupled with even more rapid opening prolongs the exponential decay (Fig. 2).

Neurotransmitter-binding site

The two ACh-binding sites of the AChR are formed at interfaces between an α -subunit and an adjacent ϵ - or δ -subunit. The α -subunit is known as the principal face and contributes three loops from discon-

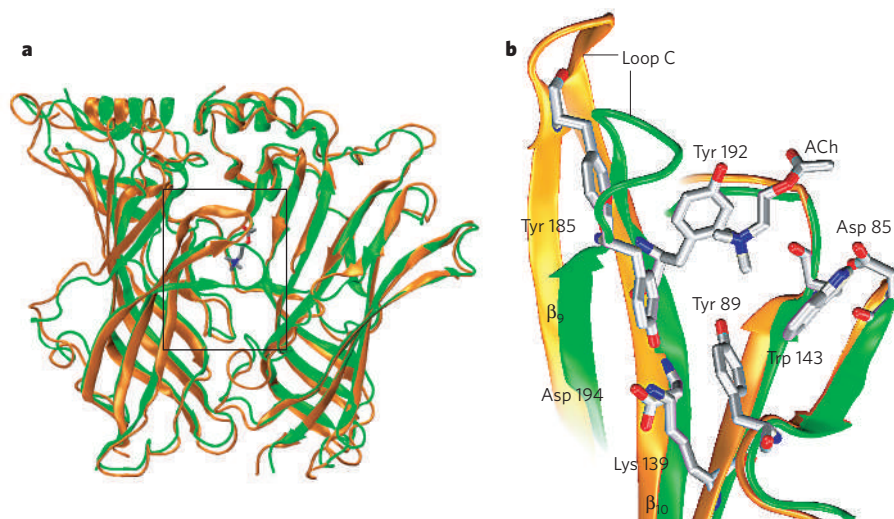


Figure 3 | Molecular dynamics simulation of AChBP showing mobility of loop C.

a, Superimposed structures of AChBP with (green) or without ACh (orange) bound after 45 ns of simulation³⁰. **b**, Close-up view of one subunit from the region boxed on the left, with key residues and bound ACh rendered in stick representation. The subunit is rotated slightly for a better view of the structural changes. Notice that in the absence of ACh, loop C is in an open, uncapped conformation. Capping by loop C brings Tyr 185 into register with Lys 139.

tinuous sections of the primary sequence, whereas the non- α -subunit is called the complementary face and contributes four discontinuous loops¹⁸. The presence of different non- α -subunits confers different affinities upon the two binding sites for agonists and competitive antagonists. Low affinity of one binding site for ACh allows rapid termination of the postsynaptic response, whereas high affinity of the other site might act as a priming mechanism²⁰. Non-equivalent affinities of the two binding sites result from intrinsic structural differences rather than from induced fit by the agonist, and cooperative interactions between the sites have not been detected.

The enrichment of aromatic and hydrophobic residues at the ACh-binding site was first revealed by mutational analyses and site-directed labelling¹⁸, and later confirmed by the structural determination of AChBP⁶ and the *Torpedo* receptor¹⁴. Five aromatic residues, α Tyr 93, α Trp 149, α Tyr 190, α Tyr 198 and ϵ Trp 55/ δ Trp 57, are highly conserved and congregate at the binding site interface. The buried nature of the interface leads to a preference for hydrophobic residues, but although these residues are not conserved, they contribute to ligand affinity; residue differences between ϵ - and δ -subunits allow ligands to select between the two binding sites¹⁸.

Crystal structures of *Lymnaea* AChBP with bound nicotine or carbamylcholine²⁶, and more recently of *Aplysia* AChBP with bound lobeline or epibatidine²⁷, show that the agonists are fully enveloped by the protein. Stabilization forces include π -cation, dipole-cation, hydrogen bonding and van der Waals interactions. At the centre of each complex is Trp 143 from the principal face of the subunit, which makes strong π -cation interactions with the agonist²⁸ (Fig. 3). The other four aromatic residues form the remainder of the binding cavity, directing their π electrons or hydroxyl groups towards the agonist. The invariant Asp 85 situated behind the central Trp polarizes its main-chain carbonyl group²⁶, further stabilizing the agonist, while in the receptor, the Asp equivalent Asp 85 hydrogen bonds to the main chain to optimize the Trp for interaction with agonist²⁹. Hydrogen bonds also form between polar moieties of the agonists and the protein main or side chains from both faces of the binding site, and in some cases include a bridging water molecule. Hydrophobic residues from both faces further stabilize agonists through van der Waals contacts. Finally, a peripheral loop at the principal face, known as loop C, caps the entrance to the binding cavity, trapping the agonists.

Mechanistic analyses predict that the agonist binds with higher affinity when the channel is open. The physical counterpart of this affinity increase is emerging from molecular dynamics (MD) simulations, crystal structures of AChBP and measurements of intrinsic Trp fluorescence. Starting with the crystal structure of AChBP, MD simulation reveals a time-dependent change of loop C from a capped to an uncapped conformation, which is prevented by bound ACh³⁰ (Fig. 3). This uncapping of loop C was also observed in MD simulation of a homology model of the homomeric $\alpha 7$ nicotinic receptor¹³. The extent of uncapping is

substantial, and would allow free diffusion of agonist and larger peptide toxins to the centre of the binding site. Recent agonist-free structures of the *Torpedo* receptor at 4 Å resolution¹⁴ and *Aplysia* AChBP at 2 Å resolution²⁷ also reveal uncapped conformations of loop C. The capped conformation with bound agonist suggests that binding flips loop C from uncapped to capped. Measurements of intrinsic Trp fluorescence of AChBP in solution show that agonist reduces accessibility of the aromatic cavity to a diffusible fluorescence quencher, presumably by promoting capping of loop C and restricting access to the cavity³⁰.

Coupling of agonist binding to the channel

The kinetics of single-channel currents at steady state^{23–25} and of macroscopic currents after rapid application of ACh^{31,32} show that binding of agonist is conveyed to the channel within tens of microseconds and with an efficiency approaching unity. The speed and efficiency of the communication are remarkable given the ~50 Å separating the binding site and channel, and implies an exquisite structural interplay of moving parts. One might expect communication between the binding site and channel to begin with structures local to the binding site, and loop C is a good candidate because its conformation changes in an agonist-dependent manner. In the uncapped conformation of loop C generated by MD simulation of AChBP³⁰, the conserved Tyr 185 (equivalent to α Tyr 190 in the receptor) is 8 Å away from the conserved Lys 139 (equivalent to α Lys 145) in β -strand 7, which forms a salt bridge with the conserved Asp 194 (equivalent to α Asp 200) in β -strand 10 (Fig. 3). However, in agonist-bound crystal structures of AChBP²⁶, loop C tilts inwards, with Tyr 185 within 2–3 Å of Lys 139, and Asp 194 displaced from the Tyr/Lys pair. In the receptor, mutation of any of these three residues markedly impairs channel gating, and all three are interdependent in contributing to gating³³. So, when agonist binds, capping of loop C draws a conserved Tyr into register with a conserved Lys, forming a hydrogen bond or possibly a salt bridge, weakening the Lys–Asp electrostatic link. The resulting inter-residue displacements could propagate to the channel through β -strand 7, which forms part of the Cys-loop, and via β -strand 10, which connects directly to M1.

At the binding–channel domain interface, β -sheet secondary structure of the binding domain meets α -helical secondary structure of the channel. This structural transition zone is a natural site of communication between binding and channel domains, and has attracted widespread attention^{12,34–36}. Composed entirely of loops spanning secondary structural elements, the binding–channel interface comprises three loops from the binding domain, the $\beta 1$ – $\beta 2$, Cys- and $\beta 8$ – $\beta 9$ loops, and one from the channel domain, the M2–M3 loop. Also present is the covalent link between β -strand 10 and M1. Communication between the two domains might be expected to require an intricate interplay between these loops.

A structural interplay between loops was demonstrated by generating a chimaeric receptor in which AChBP was substituted for the binding

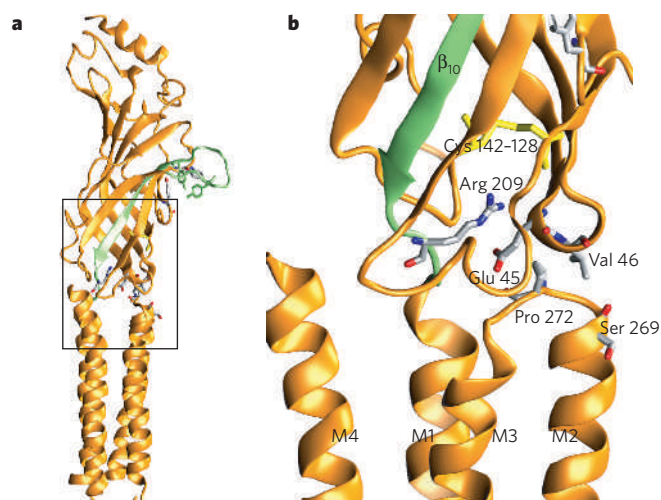


Figure 4 | Principal pathway coupling acetylcholine binding to channel gating. **a**, Side view of an α -subunit from the *Torpedo* acetylcholine receptor (Protein Data Bank code 2BG9) with loop C and β -strand 10 highlighted in green to illustrate connectivity between the binding site and the top of the channel. **b**, Close-up view of the binding-channel interface with key residues of the principal pathway rendered in stick representation. The signature Cys-loop is formed by Cys 142 and Cys 128 through a disulphide bond (yellow). The β 1– β 2 linker harbours Val 46 and Glu 45, and the M2–M3 linker harbours Ser 269 and Pro 272.

domain of the serotonin type 3A (5-HT_{3A}) receptor¹². The chimaeric receptor accumulated in robust quantities on the cell surface and bound ACh, but failed to convert binding of ACh into opening of the channel. So, although AChBP evolved from an ancestral receptor-coupled channel, structural constraints by the channel were lost, rendering loops from the binding domain structurally incompatible with the M2–M3 loop from the channel. However, substitution of residues from the 5-HT_{3A} receptor into the β 1– β 2, Cys- and β 8– β 9 loops of the AChBP/5-HT_{3A} chimaera restored structural compatibility with the M2–M3 loop, and ACh triggered opening of the channel.

That multiple loops couple binding to gating concurred with mutational analyses showing that residues in the β 1– β 2, Cys-, β 8– β 9 and M2–M3 loops and β -strand 10 contribute to gating^{12,34–40}. However, the multiplicity of loops obscured a common pathway that functionally couples the binding and channel domains.

Prospects for identifying a common pathway improved with the structural model of the *Torpedo* receptor at 4 Å resolution¹⁴. The separation and relative orientation of binding and pore domains differed from those in homology models generated from separate binding and pore domains¹², allowing fresh and more realistic leads into atomic-scale contacts that couple binding to gating. Inspection of the structure revealed a molecular pathway from the binding site to the top of the channel-forming M2 domain (Fig. 4). The pathway begins with loop C and its constituent β -strand 10, which at its distal end harbours Arg 209. The side chain of Arg 209 projects into the hydrophobic core of the subunit where it meets Glu 45 in a salt bridge. Glu 45 and the adjacent Val 46 branch from the tip of the β 1– β 2 linker and articulate with Pro 272 in the M2–M3 linker. The Val 46 side chain also inserts into a cavity created by Ser 269 to Pro 272 at the top of the channel-forming M2 helix. Capping of loop C elicited by agonist could displace Arg 209 through β -strand 10, which through electrostatic contact with Glu 45 could be transmitted to the channel. Both Arg 209 and Glu 45 are present in every member of the Cys-loop receptor family⁴¹, suggesting they form a common link between binding site and channel. Val 46, Pro 272 and Ser 269 are not conserved across the Cys-loop family of receptors, but are conserved within certain subtypes of ACh, GABA_A, glycine and 5-HT₃ receptors, suggesting that these variable residues tailor channel gating to meet a range of synaptic needs or adapt to different surrounding structures.

Mutation of any residue of the principal pathway markedly alters

channel gating, suggesting that these residues couple binding to gating⁴². Whether the residues actually couple binding to gating was tested by directly measuring the channel gating equilibrium constant using the patch clamp and determining whether juxtaposed residues are functionally interdependent. The gating equilibrium constant θ is the ratio of channel opening to closing rate constants, which are determined by analysing single-channel open and closed dwell times given a kinetic scheme in which ACh binding and channel gating are separate, sequential events^{22–25} (Fig. 2). Interdependence of juxtaposed residues was determined by applying thermodynamic mutant cycle analysis (MCA) to measurements of θ for receptors with single, double and triple mutations of key residues. MCA has been widely used to establish interdependence of residues in proteins⁴³, in some cases revealing genuine physical interactions⁴⁴, and yields a free energy associated with the degree of interdependence.

Mutations of Arg 209 and Glu 45 alter channel gating in an interdependent manner; charge-reversal of either residue impairs or abolishes channel gating, but charge-reversal of both residues restores gating to near normal⁴². Free energy of interdependence determined by MCA (Fig. 5) was similar to that for a salt bridge in the hydrophobic interior of the protein barnase⁴⁴, suggesting a direct, physical interaction. MCA also revealed strong interdependence of the physically contiguous residue triads Glu 45/Val 46/Pro 272 and Val 46/Pro 272/Ser 269, while it demonstrated independence of the spatially separate pairs Glu 45/Ser 269 and Glu 45/Val 46. Although residue interdependence does not necessarily disclose physical contact, the correlation between proximity and interdependence, and that between separation and independence, provides compelling evidence for direct physical interactions coupling binding to gating.

Further insight into how binding couples to gating emerged from the substitution of unnatural amino acids for Pro 272 in the M2–M3 loop of the 5-HT_{3A} receptor⁴⁵. Proline is unique because its side chain forms a closed loop with the protein backbone, allowing *cis* and *trans* conformations. Channel gating, inferred from changes in the EC₅₀ (the concentration of agonist that produces 50% of the maximum response) for activation by the agonist, was impaired or eliminated by amino acids that prefer the *trans* conformation, whereas gating was facilitated by amino acids that prefer the *cis* conformation. So, isomerization of Pro 272 into the *cis* conformation may bend the M2–M3 loop, which may in turn dislodge M2 into the open channel conformation.

Gating of the channel

Beyond providing an electrical conduit through the cell membrane, the channel also selects for positively or negatively charged ions and gates

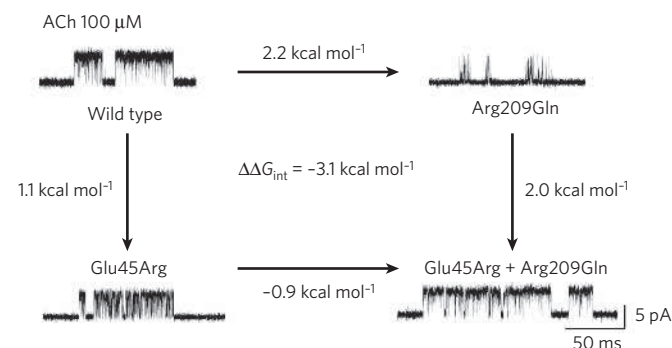


Figure 5 | Energetic coupling of a charged residue pair that links agonist binding to channel gating. Thermodynamic mutant cycle analysis of charge reversal mutations of Arg 209 and Glu 45 (ref. 42). Single-channel currents elicited by 100 μ M acetylcholine (ACh) are shown for receptors containing the indicated mutations. Free energy changes for channel gating are shown for each limb of the cycle. The overall coupling free energy, $\Delta\Delta G_{\text{int}}$, is given by $-RT \ln[(\theta_{\text{ww}}\theta_{\text{mm}})/(\theta_{\text{wm}}\theta_{\text{mw}})]$, where the θ s are gating equilibrium constants for wild type (ww, $\theta = 27$), single mutant (mw, Glu45Arg, $\theta = 4.1$, and wm, Arg209Gln; $\theta = 0.59$) and double mutant (mm, $\theta = 19$) receptors.

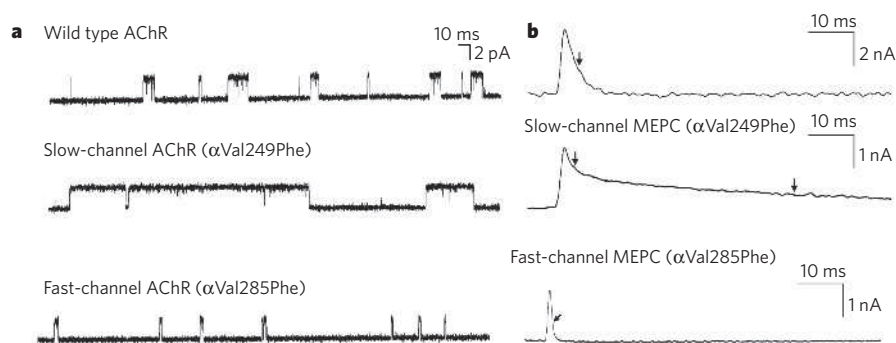


Figure 6 | Slow- and fast-channel phenotypes of congenital myasthenic syndrome. **a**, Single-channel currents from the indicated receptors expressed in human embryonic kidney fibroblasts. **b**, Miniature endplate currents (MEPC) recorded from endplates of a control subject, a patient with the α Val249Phe slow-channel mutation, and a patient with the α Val285Phe fast-channel mutation. Arrows indicate time constants of the current decay. The slow-channel phenotype exhibits prolonged single-channel currents and slow decay of the MEPC, whereas the fast-channel phenotype exhibits brief single-channel currents and accelerated decay of the MEPC. Both phenotypes cause muscle weakness but by different mechanisms (see Fig. 2).

their flow. Ion selectivity is achieved by charged residues flanking M2, at both extracellular and intracellular ends⁴⁶, and in the M3–M4 cytoplasmic domain⁴⁷. Channel gating primarily involves motion of the M2 helix⁴⁸, while the other three helices might provide a supporting scaffold.

The barrier to ion flow is a narrow constriction formed by M2 helices from the five subunits, and is called the resting gate (Fig. 1). It has been localized by the substituted cysteine accessibility method (SCAM) applied to the mouse muscle AChR⁴⁹, and by structural determination of the *Torpedo* receptor¹⁴. Both methods localize the gate to within the cytoplasmic half of the bilayer, but SCAM places it about three helical turns closer to the cytoplasm than in the *Torpedo* structure.

Gating leads to a transient removal of the barrier to ion flow. In the resting gate, the narrow constriction has a diameter of about 3 Å, which increases to about 8 Å when the gate opens. The fundamental motion underlying gating appears to be a rotation of the M2 helices, as judged from a 9-Å resolution structure of the *Torpedo* receptor subjected to brief application of ACh⁴⁸. Elastic vibration analysis of a model of the homomeric $\alpha 7$ receptor revealed a similar rotation of M2, with a major mode exhibiting twisting of the receptor about its long axis⁵⁰. On the other hand, a rocking motion of M2 about a fulcrum some distance through the membrane was suggested by experiments assessing the ability of histidine substituted at different levels of M2 to form Zn^{2+} -binding sites⁵¹ or of lysine substitutions to block the open channel in a pH-dependent manner⁵².

Channel gating, like agonist binding, contributes to the rapid rise and exponential decay of the EPC (Fig. 2). The rate constant for chan-

nel opening is high, allowing opening of the ACh-occupied channel within tens of microseconds, and also allowing repeated opening during each occupancy. The rate constant for channel closing is about 20-fold slower, yielding an open-state lifetime of around half a millisecond, and contributing to the rate of EPC decay.

Transition state for gating

So far, the mechanistic discussion has centred on stable states of the receptor, open and closed, but equally important is the transition state that connects them. The transition state is important because its free energy relative to closed and open states determines the channel gating rate constants. The structure of the transition state in a given location can be inferred from changes in opening and closing rate constants after a series of focal structural perturbations. This is done by plotting logarithms of opening rate constant against gating equilibrium constant, called a rate–equilibrium free energy relationship (REFER), which for perturbations of gating-sensitive structures gives a straight line with a slope ϕ . If ϕ approaches unity, changes in gating are dominated by changes in the opening rate constant, and the structure of the transition state in the vicinity of the perturbation resembles the closed state, whereas if ϕ approaches zero, changes in gating are dominated by changes in the closing rate constant and the structure of the transition state in the vicinity of the perturbation resembles the open state⁵³. A sweeping range of structural perturbations reveals a gradient of ϕ values from the binding site to the channel, indicating that the transition state is mainly closed at the binding site, but mainly open at the channel⁵⁴. Detailed analyses indicate that similar transition state conformations occur in structural blocks, suggesting that channel opening proceeds through a sequence of domain shifts^{35,55,56}.

Recent studies suggest that instead of forming a single barrier between closed and open states, the transition state is a raised plateau with a series of small energy barriers, or corrugations, each corresponding to the shift of a domain^{57,58}. A corrugated transition state is suggested by theoretical work showing that perturbation of a corrugation proximal to the closed state mainly affects the rate of channel opening, whereas perturbation proximal to the open state mainly affects the rate of channel closing. The corrugated model accounts for the change in gating rate versus equilibrium constants, but predicts a nonlinear rather than the observed linear relationship; the available measurements might be confined to the linear portion of the REFER plot, or the model of the transition state might require modification.

Congenital myasthenic syndromes

Insights into structure–function relationships of Cys-loop receptors have emerged from studies of naturally occurring mutations known as channelopathies, which cause a diverse range of neurological diseases. Both cationic and anionic Cys-loop receptors are targeted and include

Table 1 | Classification of the congenital myasthenic syndromes

Site of defect	Index cases
Presynaptic (7%)	
ChAT deficiency*	10
Paucity of synaptic vesicles	1
Congenital Lambert-Eaton syndrome-like	1
Other presynaptic defects	4
Synaptic space (16%)	
Endplate AChE deficiency*	35
Postsynaptic (77%)	
Primary AChR kinetic defect*	52
Primary AChR deficiency*	89
Rapsyn deficiency*	32
Na ⁺ channel myasthenia*	1
Total	226

Classification of the congenital myasthenic syndromes was based on 226 index cases: 115 patients had intercostal muscle biopsies at the Mayo Clinic; 111 patients were studied using DNA isolated from blood. AChE, acetylcholinesterase; AChR, acetylcholine receptor; ChAT, choline acetyltransferase.

*Genetic defect identified.

mutations of muscle nicotinic AChRs causing congenital myasthenic syndromes (CMSs)⁵⁹, mutations of GABA_A and neuronal nicotinic AChRs causing epilepsy⁶⁰, and mutations of glycine receptors causing hyperekplexia⁶¹. Here we focus our attention on CMSs.

Since the early 1990s, combined clinical, *in vitro* electrophysiological, ultrastructural and molecular genetic studies have uncovered molecular bases of CMSs. By the end of 2005, more than 500 kinships had been identified worldwide⁶², and the subunits of AChR⁶³, ColQ (the structural subunit of acetylcholinesterase, AChE)⁶⁴, choline acetyltransferase⁶⁵, rapsyn (receptor-associated protein of the synapse)⁶⁶, MuSK (muscle-specific receptor tyrosine kinase)⁶⁷, and the muscle voltage-gated sodium channel (Na_v1.4)⁶⁸ emerged as targets of the CMSs. Table 1 classifies the CMSs according to the site of the defect and relative frequency in the Mayo cohort of 226 kinships. Notably, more than 75% of CMSs are postsynaptic. Most postsynaptic mutations reside in AChR, where they disrupt its activation kinetics or curtail expression⁶³.

Low-expressor mutations in AChR

CMSs with severe motor endplate AChR deficiency result from different homozygous or more frequently from heterozygous recessive mutations in AChR subunit genes. These mutations concentrate in the ϵ -subunit and are the commonest cause of CMS. The likely reason is that expression of the fetal γ -subunit, although low, partly compensates for absence of the ϵ -subunit, whereas patients harbouring null mutations in non- ϵ -subunit genes might not survive for lack of a substituting subunit⁶⁹.

Defects in AChR kinetics

The slow-channel CMS is caused by dominant gain-of-function mutations, and the fast-channel syndrome by recessive loss-of-function mutations. The two syndromes are physiological opposites, with the

slow-channel syndrome showing prolonged decay of the EPC, and the fast-channel syndrome showing accelerated decay (Fig. 6). The two syndromes also respond differently to therapy. The slow-channel syndrome improves with the long-lived open-channel blockers quinidine^{70,71} and fluoxetine⁷², which attenuate the synaptic response to ACh. The fast-channel syndrome improves with 3,4-diaminopyridine, which increases the number of quanta released by a nerve stimulus, and AChE inhibitors, which augment the synaptic response²¹.

No fewer than 18 slow-channel mutations have been identified so far. Almost all occur in the ϵ - and α -subunits, appearing in the extracellular domain or in M1 or M2 (ref. 63). Because the prolonged synaptic potential outlasts the refractory period of the muscle fibre AP, single nerve stimuli evoke more than one AP, and summation of synaptic potentials at physiological rates of stimulation causes depolarization block of the endplate. The prolonged as well as spontaneous channel openings overload the postsynaptic region with cations, including Ca²⁺, causing an endplate myopathy manifested by the degeneration of junctional folds, loss of AChR from the folds, widening of the synaptic space, and vacuolar change and apoptosis of the subsynaptic nuclei^{73–75}.

Thirteen fast-channel mutations have been documented to date. These appeared in the extracellular domains of α -, δ - and ϵ -subunits, in transmembrane domains of the α - and δ -subunits, and in the M3–M4 loop of the ϵ -subunit. In most cases, a null mutation is present but in a different allele from the one harbouring the fast-channel mutation so that the phenotype is determined by the fast-channel mutation.

Mutations affecting agonist binding

The fast-channel mutation ϵ Asp175Asn at the entrance to the binding site slows ACh association and speeds its dissociation, probably by nullifying electrostatic attraction of the cationic ACh towards the binding cavity (Fig. 7); only the α - ϵ site is affected, leaving the α - δ site unaltered⁷⁶. The slow-channel mutation α Gly153Ser near the centre of the binding site has the opposite effect, speeding ACh association and slowing dissociation, which are probably secondary to changes in the protein backbone harbouring the nearby α Trp 149, which coordinates ACh⁷⁷. Neither ϵ Asp175Asn nor α Gly153Ser is state specific; each affects ACh binding to open and closed states equally.

But CMS can arise from state-selective effects on ACh binding, as in the fast-channel mutation ϵ Pro121Leu deep in the binding cavity (Fig. 7), which diminishes affinity for the open state with little effect on the closed state, profoundly diminishing gating efficiency²¹. A likely explanation is that ϵ Pro 121 is required for local rearrangements that increase ACh affinity for the open state, but is not essential to maintain low affinity for the closed state. Yet another twist occurs with the fast-channel mutation ϵ Asn182Tyr at the entrance of the binding site, which reverses the order in which ACh occupies the two binding sites⁷⁶. In the closed state, the α - ϵ site normally binds the agonist with low affinity, and the α - δ site binds with high affinity, but ϵ Asn182Tyr speeds ACh association and slows dissociation, increasing the affinity of the α - ϵ site and reversing the order of occupancy. High affinity of both binding sites in the closed state reduces the relative preference of ACh for the open over the closed state, diminishing gating efficiency.

The binding site can also be altered from a distance, the so-called allosteric affect. Some allosteric mutations alter ACh binding with minor effects on channel gating, such as the slow-channel mutation α Asn217Lys in the M1 domain⁷⁸, and the fast-channel mutation α Val132Leu at the interface between binding and channel domains⁷⁹ (Fig. 7). The two mutations have opposite effects on closed state affinity, selectively affecting association (α Val132Leu) or dissociation (α Asn217Lys), indicating that the binding site in the closed state is structurally altered. In both cases a contiguous structure links the site of the mutation to the binding site; M1 and the contiguous β -strand 10 link α Asn 217 to loop C, whereas β -strand 7 links α Val 132 to the central α Trp 149. So, residues far from the binding site allosterically maintain the low affinity of the resting state. Other mutations that allosterically affect ACh binding include the slow-channel mutations ϵ Leu78Pro in β -strand 3 (ref. 80) and ϵ Leu221Phe in M1 (ref. 81).

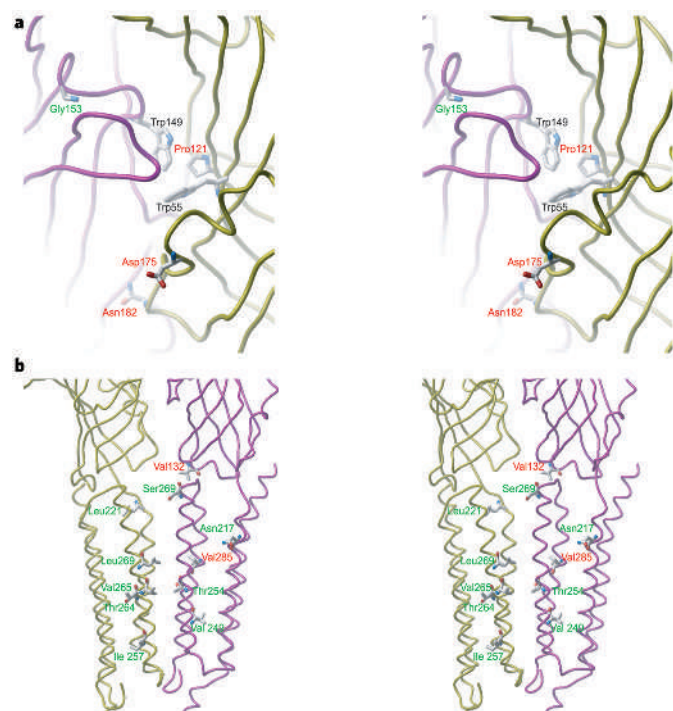


Figure 7 | CMS mutations occur in both binding and channel domains of the AChR. Residues mutated in CMS are mapped on a structural model of the human AChR, which was generated using the *Torpedo* receptor structural model as the template (Protein Data Bank code 2BG9). **a**, Stereo image of the binding site viewed from the periphery with residues mutated in CMS shown in stick representation. Residues α Trp 149 and ϵ Trp 55 are also shown to provide a frame of reference. The α -subunit is highlighted in magenta and the ϵ -subunit in gold. Sites of fast-channel CMS are indicated by red type, whereas those of slow-channel CMS are indicated in green type. **b**, Stereo image of the channel domain viewed from the central vestibule, with residues mutated in CMS shown in stick representation.

Some allosteric mutations affect both agonist binding and channel gating. The slow-channel mutation α Val249Phe, near the narrow constriction formed by M2 (Fig. 7), enhances channel gating and increases affinity of the closed state, allowing nanomolar concentrations of ACh to elicit long-lived channel events⁷⁴. Affinity increases at only one of the two binding sites, even though the mutation is present in both α -subunits, suggesting that the M2 helices contribute unequally to resting state affinity. That a mutation in M2 affects the remote binding site underscores the bi-directional nature of communication in receptor-coupled channels.

CMS mutations that alter ACh binding affect the amplitude and time course of the EPC. Mutations that enhance binding to the closed state slow the rate of ACh dissociation, allowing multiple re-openings during ACh occupancy and prolonging the EPC (Fig. 2). Mutations that diminish binding to the closed state speed the rate of ACh dissociation, allowing the channel to open only once per ACh occupancy, accelerating the EPC decay. The amplitude of the EPC is determined by the fraction of channels that open, given by (opening rate)/(opening rate + ACh dissociation rate), which is normally in the range 0.82–0.96 (refs 21, 81), so that increasing the dissociation rate relative to the channel opening rate reduces the peak of the EPC.

Mutations affecting channel gating

The receptor channel opens in the absence of agonist but at a very low frequency. Thus, slow-channel mutations that enhance ACh-mediated gating might be expected to alter this intrinsic gating step. The first slow-channel mutation defined at the molecular level localizes to the M2 helix of the ϵ -subunit (ϵ Thr264Pro; Fig. 7), and dramatically increases spontaneous opening of the channel⁷³. Since then a host of slow-channel mutations has been discovered in M2 of different subunits^{59,63}, and many increase spontaneous opening. So, structural perturbation of M2 itself can tip the balance between closed and open, allowing binding of ACh to open the channel with greater efficiency.

Changes in gating efficiency are directly assessed by measuring the channel opening and closing rate constants. Which step is affected the most depends on the conformation of the mutant residue in the transition state separating closed and open states. The more open-like transition state characteristic of M2, revealed by REFER analysis, predicts a predominant slowing of channel closing and a minimal speeding of channel opening⁵⁴; slow-channel mutations in M2 invariably exhibit a marked slowing of the rate of channel closing.

In contrast, mutations in the channel can impair channel gating efficiency and cause a fast-channel syndrome. One such mutation, α Val285Ile in the M3 helix (Fig. 7), hints at the nature of conformational changes of M2 during gating⁸². The space separating M3 and M2 widens from intracellular to extracellular boundaries, and α Val 285 projects into this space at a level about midway between boundaries. Gating free energy was found to scale linearly with the size of the moiety attached to the C β of residue α 285, with a smaller moiety enhancing gating and a larger moiety impairing gating. Because gating probably involves rigid body motion of M2, whether it be rotation or tilting, mobility of M2 within this space is crucial, and the larger side chain in α Val285Ile restricts M2, impairing gating.

CMS mutations that enhance channel gating maintain a rapid rise of the EPC, owing to rapid channel opening, but by slowing channel closing they slow the EPC decay (Fig. 2). Rapid channel opening further slows the EPC decay by promoting reopening of the channel before ACh can dissociate, but minimally affects EPC amplitude. In contrast, mutations that impair gating slow the rise of the EPC, owing to slow-channel opening, but by speeding channel closing they accelerate the decay of the EPC. Decay of the EPC accelerates further because slow-channel opening allows ACh to dissociate before reopening can occur. EPC amplitude is reduced because channel opening slows relative to ACh dissociation.

Future prospects

Now that the atomic structural age has dawned on the field of Cys-loop receptors, structure–function relationships will be established at an increased pace. Furthermore, computational predictions of global

conformational changes that underlie activation of the receptor are within reach. These predictions can be tested by structural perturbations coupled with single-receptor functional analyses, lending deeper insight into structure–function relationships. Ultimately, complete atomic structures at higher resolution and in closed and open states will be needed. About a 100-fold increase over current computational speed should then allow simulation of a cycle of receptor activation, starting with the atomic structure. The consequences of disease-causing mutations might then become predictable, and designer drugs developed to target specific sites on receptors and to distinguish between different receptor subtypes. ■

1. Takeuchi, A. & Takeuchi, N. Active phase of frog's endplate potential. *J. Neurophysiol.* **22**, 395–412 (1959).
2. Montal, M. Reconstitution of channel proteins from excitable cells in planar lipid bilayer membranes. *J. Membr. Biol.* **98**, 101–115 (1989).
3. Unwin, N. Nicotinic acetylcholine receptor at 9 Å resolution. *J. Mol. Biol.* **229**, 1101–1124 (1993).
4. Miyazawa, A., Fujiyoshi, Y. & Unwin, N. Nicotinic acetylcholine receptor at 4.6 Å resolution: transverse tunnels in the channel wall. *J. Mol. Biol.* **288**, 765–786 (1999).
5. Smit, A. B. *et al.* A *glia*-derived acetylcholine-binding protein that modulates synaptic transmission. *Nature* **411**, 261–268 (2001).
6. Brejck, K. *et al.* Crystal structure of ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. *Nature* **411**, 269–276 (2001).
7. Sine, S. M., Wang, H.-L. & Bren, N. Lysine scanning mutagenesis delineates structure of nicotinic receptor binding domain. *J. Biol. Chem.* **277**, 29210–29223 (2002).
8. Le Novère, N., Grutter, T. & Changeux, J. P. Models of the extracellular domain of the nicotinic receptors and of agonist- and Ca²⁺-binding sites. *Proc. Natl Acad. Sci. USA* **99**, 3210–3215 (2002).
9. Molles, B. E., Tsigelny, I., Gao, S. X., Sine, S. M. & Taylor, P. Residues in the nicotinic acetylcholine receptor interact to confer selectivity of α -bungarotoxin for the α - ϵ subunit interface site. *Biochemistry* **41**, 7895–7906 (2002).
10. Schapira, M., Abagyan, R. & Totrov, M. Structural model of nicotinic receptor isoforms bound to acetylcholine and nicotine. *BMC Struct. Biol.* **2**, 1–8 (2002).
11. Miyazawa, A., Fujiyoshi, Y. & Unwin, N. Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **424**, 949–955 (2003).
12. Bouzat, C. *et al.* Coupling of agonist binding to channel gating in an ACh-binding protein linked to an ion channel. *Nature* **430**, 896–900 (2004).
13. Law, R. J., Henchman, R. H. & McCammon, J. A. A gating mechanism proposed from a simulation of a human $\alpha 7$ nicotinic acetylcholine receptor. *Proc. Natl Acad. Sci. USA* **102**, 6813–6818 (2005).
14. Unwin, N. Refined structure of the nicotinic acetylcholine receptor at 4 Å resolution. *J. Mol. Biol.* **346**, 967–989 (2005).
15. Lingle, C. J., Maconochie, D. & Steinbach, J. H. Activation of skeletal muscle nicotinic acetylcholine receptors. *J. Membr. Biol.* **126**, 195–217 (1992).
16. Edmonds, B., Gibb, A. J. & Colquhoun, D. Mechanisms of activation of muscle nicotinic acetylcholine receptors and the time course of endplate currents. *Annu. Rev. Physiol.* **57**, 469–493 (1995).
17. Jackson, M. B. Ligand-gated channel: postsynaptic receptors and drug targets. *Adv. Neurol.* **79**, 511–524 (1999).
18. Sine, S. M. The nicotinic receptor ligand binding domain. *J. Neurobiol.* **53**, 431–446 (2002).
19. Jackson, M. B. Kinetics of unliganded acetylcholine receptor channel gating. *Biophys. J.* **49**, 663–672 (1986).
20. Jackson, M. B. Perfection of a synaptic receptor: kinetics and energetics of the acetylcholine receptor. *Proc. Natl Acad. Sci. USA* **86**, 2199–2203 (1989).
21. Ohno, K. *et al.* Congenital myasthenic syndrome caused by decreased agonist binding affinity due to a mutation in the acetylcholine receptor ϵ subunit. *Neuron* **17**, 157–170 (1996).
22. Jackson, M. B. Dependence of acetylcholine receptor kinetics on agonist concentration in cultured mouse muscle fibres. *J. Physiol. (Lond.)* **397**, 555–583 (1988).
23. Sine, S. M., Claudio, T. & Sigworth, F. J. Activation of *Torpedo* acetylcholine receptors expressed in mouse fibroblasts: single-channel current kinetics reveal distinct agonist binding affinities. *J. Gen. Physiol.* **96**, 395–437 (1990).
24. Zhang, Y., Chen, J. & Auerbach, A. Activation of recombinant mouse acetylcholine receptors by acetylcholine, carbamylcholine and tetraethylammonium. *J. Physiol. (Lond.)* **486**, 189–206 (1995).
25. Colquhoun, D. & Sakmann, B. Fluctuations in the microsecond time range of the current through single acetylcholine receptor ion channels. *Nature* **294**, 464–466 (1981).
26. Celie, P. H. N. *et al.* Nicotine and carbamylcholine binding to nicotinic acetylcholine receptors as studied in AChBP crystal structures. *Neuron* **41**, 907–914 (2004).
27. Hansen, S. B. *et al.* Structures of *Aplysia* AChBP complexes with nicotinic agonists and antagonists reveal distinctive binding interfaces and conformations. *EMBO J.* **24**, 3635–3646 (2005).
28. Zhong, W. *et al.* From *ab initio* quantum mechanics to molecular neurobiology: a cation- π binding site in the nicotinic receptor. *Proc. Natl Acad. Sci. USA* **95**, 12088–12093 (1998).
29. Lee, W. Y. & Sine, S. M. Invariant aspartic acid in muscle nicotinic receptor contributes selectively to the kinetics of agonist binding. *J. Gen. Physiol.* **124**, 555–567 (2004).
30. Gao, F. *et al.* Agonist-mediated conformational changes in acetylcholine-binding protein revealed by simulation and intrinsic tryptophan fluorescence. *J. Biol. Chem.* **280**, 8443–8451 (2005).
31. Liu, Y. & Dilger, J. P. Opening rate of acetylcholine receptor channels. *Biophys. J.* **60**, 424–432 (1991).

32. Maconochie, D. & Steinbach, J. H. The channel opening rate of adult- and fetal-type mouse muscle nicotinic receptors activated by acetylcholine. *J. Physiol. (Lond.)* **506**, 53–72 (1998).
33. Mukhtasimova, F., Free, C. & Sine, S. M. Initial coupling of binding to gating mediated by conserved residues in muscle nicotinic receptor. *J. Gen. Physiol.* **506**, 53–72 (2005).
34. Kash, T., Jenkins, A., Kelly, J., Trudell, J. & Harrison, N. L. Coupling of agonist binding to channel gating in the GABA_A receptor. *Nature* **421**, 272–275 (2003).
35. Chakrapani, S., Bailey, T. D. & Auerbach, A. Gating dynamics of the acetylcholine receptor extracellular domain. *J. Gen. Physiol.* **123**, 341–356 (2004).
36. Xiu, X., Hanek, A. P., Wang, J., Lester, H. A. & Dougherty, D. A. A unified view of the role of electrostatic interactions in modulating the gating of Cys loop receptors. *J. Biol. Chem.* **280**, 41655–41666 (2005).
37. Lyford, L. K., Sproul, A. D., Eddins, D., McLaughlin, J. T. & Rosenberg, R. L. Agonist-induced conformational changes in the extracellular domain of $\alpha 7$ nicotinic acetylcholine receptors. *Mol. Pharmacol.* **64**, 650–658 (2003).
38. Schofield, C. M., Jenkins, A. & Harrison, N. L. A highly conserved aspartic acid residue in the signature disulfide loop of the $\alpha 1$ subunit is a determinant of gating in the glycine receptor. *J. Biol. Chem.* **278**, 34079–34083 (2003).
39. Absalom, N. L., Lewis, T. M., Kaplan, W., Pierce, K. D. & Schofield, P. R. Role of charged residues in coupling ligand binding and channel activation in the extracellular domain of the glycine receptor. *J. Biol. Chem.* **278**, 50151–50157 (2003).
40. Sala, F., Mulet, J., Sala, S., Gerber, S. & Criado, M. Charged amino acids of the N-terminal domain are involved in coupling binding and gating in $\alpha 7$ nicotinic receptors. *J. Biol. Chem.* **280**, 6642–6647 (2005).
41. Le Novère, N. & Changeux, J.-P. The Ligand-Gated Ion Channel database: an example of a sequence database in neuroscience. *Phil. Trans. R. Soc. Lond. B* **356**, 1121–1130 (2001).
42. Lee, W. Y. & Sine, S. M. Principal pathway coupling agonist binding to channel gating in nicotinic receptors. *Nature* **438**, 243–247 (2005).
43. Serrano, L., Horovitz, A., Avron, B., Bycroft, M. & Fersht, A. Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry* **29**, 9343–9352 (1990).
44. Vaughan, C. K., Harryson, P., Buckel, A. M. & Fersht, A. R. A structural double-mutant cycle: estimating the strength of a buried salt bridge in barnase. *Acta Crystallogr. D* **58**, 591–600 (2002).
45. Lummis, S. C. R. *et al.* *Cis-trans* isomerization at a proline opens the pore of a neurotransmitter-gated ion channel. *Nature* **438**, 248–252 (2005).
46. Imoto, K. *et al.* Rings of negatively charged amino acids determine the acetylcholine receptor channel conductance. *Nature* **335**, 645–648 (1988).
47. Kelley, S. P., Dunlop, J. I., Kirness, E. F., Lambert, J. J. & Peters, J. A. A cytoplasmic region determines single-channel conductance in 5-HT₃ receptors. *Nature* **424**, 321–324 (2003).
48. Unwin, N. Acetylcholine receptor channel imaged in the open state. *Nature* **373**, 37–43 (1995).
49. Wilson, G. G. & Karlin, A. The location of the gate in the acetylcholine receptor channel. *Neuron* **20**, 1269–1281 (1998).
50. Taly, A. *et al.* Normal mode analysis suggests a quaternary twist model for the nicotinic receptor gating mechanism. *Biophys. J.* **88**, 3954–3965 (2005).
51. Paas, Y. *et al.* Pore conformations and gating mechanism of a Cys-loop receptor. *Proc. Natl Acad. Sci. USA* **102**, 15877–15882 (2005).
52. Cymes, G. D., Ni, Y. & Grosman, C. Probing ion-channel pores one proton at a time. *Nature* **438**, 975–980 (2005).
53. Fersht, A. R., Leatherbarrow, R. J. & Wells, T. N. C. Structure–activity relationships in engineered proteins: analysis of use of binding energy by linear free-energy relationships. *Biochemistry* **26**, 630–638 (1987).
54. Grosman, C., Zhou, M. & Auerbach, A. Mapping the conformational wave of acetylcholine receptor channel gating. *Nature* **403**, 773–776 (2002).
55. Cymes, G. D., Grosman, C. & Auerbach, A. Structure of the transition state of gating in the acetylcholine receptor channel pore: a Φ -value analysis. *Biochemistry* **41**, 5548–5555 (2002).
56. Chakrapani, S., Bailey, T. D. & Auerbach, A. The role of loop 5 in acetylcholine receptor channel gating. *J. Gen. Physiol.* **122**, 521–539 (2003).
57. Auerbach, A. Gating of acetylcholine receptor channels: brownian motion across a broad transition state. *Proc. Natl Acad. Sci. USA* **102**, 1408–1412 (2005).
58. Zhou, Y., Pearson, J. E. & Auerbach, A. Φ -value analysis of a linear, sequential reaction mechanism: theory and application to channel gating. *Biophys. J.* **89**, 3680–3685 (2005).
59. Engel, A. G., Ohno, K. & Sine, S. M. Sleuthing molecular targets for neurological diseases at the neuromuscular junction. *Nature Rev. Neurosci.* **4**, 339–352 (2003).
60. Steinlein, O. K. Genetic mechanisms that underlie epilepsy. *Nature Rev. Neurosci.* **5**, 400–408 (2004).
61. Shiang, R. *et al.* Mutational analysis of familial and sporadic hyperekplexia. *Ann. Neurol.* **38**, 85–91 (1995).
62. Beeson, D., Hantai, D., Lochmuller, H. & Engel, A. G. 126th International Workshop: congenital myasthenic syndromes. *Neuromuscul. Disord.* **15**, 498–512 (2005).
63. Engel, A. G. & Sine, S. M. Current understanding of congenital myasthenic syndromes. *Curr. Opin. Pharmacol.* **5**, 306–321 (2005).
64. Ohno, K., Brengman, J. M., Tsujino, A. & Engel, A. G. Human endplate acetylcholinesterase deficiency caused by mutations in the collagen-like tail subunit (ColQ) of the asymmetric enzyme. *Proc. Natl Acad. Sci. USA* **95**, 9654–9659 (1998).
65. Ohno, K. *et al.* Choline acetyltransferase mutations cause myasthenic syndrome associated with episodic apnea in humans. *Proc. Natl Acad. Sci. USA* **98**, 2017–2022 (2001).
66. Ohno, K. *et al.* Rapsyn mutations in humans cause endplate acetylcholine receptor deficiency and myasthenic syndrome. *Am. J. Hum. Genet.* **70**, 875–885 (2002).
67. Chevessier, F. *et al.* MUSK, a new target for mutations causing congenital myasthenic syndrome. *Hum. Mol. Genet.* **13**, 3229–3240 (2004).
68. Tsujino, A. *et al.* Myasthenic syndrome caused by mutation of the *SCN4A* sodium channel. *Proc. Natl Acad. Sci. USA* **100**, 7377–7382 (2003).
69. Ohno, K. *et al.* Congenital myasthenic syndromes due to heteroallelic nonsense/missense mutations in the acetylcholine receptor ϵ subunit gene: identification and functional characterization of six new mutations. *Hum. Mol. Genet.* **6**, 753–766 (1997).
70. Fukudome, T., Ohno, K., Brengman, J. M. & Engel, A. G. Quinidine normalizes the open duration of slow-channel mutants of the acetylcholine receptor. *Neuroreport* **9**, 1907–1911 (1998).
71. Harper, C. M. & Engel, A. G. Quinidine sulfate therapy for the slow-channel congenital myasthenic syndrome. *Ann. Neurol.* **43**, 480–484 (1998).
72. Harper, C. M., Fukudome, T. & Engel, A. G. Treatment of slow channel congenital myasthenic syndrome with fluoxetine. *Neurology* **60**, 170–173 (2003).
73. Ohno, K. *et al.* Congenital myasthenic syndrome caused by prolonged acetylcholine receptor channel openings due to a mutation in the M2 domain of the ϵ subunit. *Proc. Natl Acad. Sci. USA* **92**, 758–762 (1995).
74. Milone, M. *et al.* Slow-channel syndrome caused by enhanced activation, desensitization, and agonist binding affinity due to mutation in the M2 domain of the acetylcholine receptor α subunit. *J. Neurosci.* **17**, 5651–5665 (1997).
75. Engel, A. G. *et al.* New mutations in acetylcholine receptor subunit genes reveal heterogeneity in the slow-channel congenital myasthenic syndrome. *Hum. Mol. Genet.* **5**, 1217–1227 (1996).
76. Sine, S. M. *et al.* Naturally occurring mutations at the acetylcholine receptor binding site independently alter ACh binding and channel gating. *J. Gen. Physiol.* **120**, 483–496 (2002).
77. Sine, S. M. *et al.* Mutation of the acetylcholine receptor α subunit causes a slow-channel myasthenic syndrome by enhancing agonist binding affinity. *Neuron* **15**, 229–239 (1995).
78. Wang, H.-L. *et al.* Mutation in the M1 domain of the acetylcholine receptor α subunit decreases the rate of agonist dissociation. *J. Gen. Physiol.* **109**, 757–766 (1997).
79. Shen, X.-M. *et al.* Mutation causing severe myasthenia reveals functional asymmetry of AChR signature Cys-loops in agonist binding and gating. *J. Clin. Invest.* **111**, 497–505 (2003).
80. Shelley, C. & Colquhoun, D. A human congenital myasthenia-causing mutation (ϵ L78P) of the muscle nicotinic acetylcholine receptor with unusual single channel properties. *J. Physiol. (Lond.)* **564**, 377–396 (2005).
81. Hatton, C. J., Shelley, C., Brydson, M., Beeson, D. & Colquhoun, D. Properties of the human muscle nicotinic receptor, and of the slow-channel myasthenic syndrome mutant ϵ L221F, inferred from maximum likelihood fits. *J. Physiol. (Lond.)* **547**, 729–760 (2003).
82. Wang, H.-L. *et al.* Acetylcholine receptor M3 domain: stereochemical and volume contributions to channel gating. *Nature Neurosci.* **2**, 226–233 (1999).

Acknowledgements We thank H.-L. Wang and W. Y. Lee for generating the structural figures. Work in the authors' laboratories was supported by grants to S.M.S. and A.G.E. from the National Institutes of Health, and a grant to A.G.E. from the Muscular Dystrophy Association.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.M.S. (sine@mayo.edu).

Glutamate receptors at atomic resolution

Mark L. Mayer¹

At synapses throughout the brain and spinal cord, the amino-acid glutamate is the major excitatory neurotransmitter. During evolution, a family of glutamate-receptor ion channels seems to have been assembled from a kit consisting of discrete ligand-binding, ion-channel, modulatory and cytoplasmic domains. Crystallographic studies that exploit this unique architecture have greatly aided structural analysis of the ligand-binding core, but the results also pose a formidable challenge, namely that of resolving the allosteric mechanisms by which individual domains communicate and function in an intact receptor.

The glutamate-receptor ion channels (iGluRs), namely AMPA, kainate and NMDA receptors, are the major mediators of excitatory synaptic transmission in the central nervous system^{1–3}. Each subtype has a unique role. NMDA receptors, which have attracted enormous interest, act as coincidence detectors. These monitor changes in membrane potential and the presence of glutamate in the synaptic cleft⁴. AMPA receptors have key roles in mediating the rapid excitatory synaptic current, the kinetics of which is tuned over about a fivefold range in different structures of the brain⁵. Kainate receptors have a major modulatory role at both presynaptic and postsynaptic sites, and kainate-receptor agonists have long been known to be potent convulsants and environmental toxins in the food supply^{6,7}. These receptors were identified through traditional pharmacological approaches, and their naming reflects this: AMPA, kainate and NMDA are synthetic agonists and naturally occurring toxins not found in the brain⁸.

AMPA receptors are encoded by four genes (*GluR1–4*), kainate receptors by three genes for 'low-affinity' subunits (*GluR5–7*) and two for 'high-affinity' subunits (*KA1* and *KA2*), and NMDA receptors by seven genes (*NR1*, *NR2A–D*, *NR3A* and *NR3B*). A fourth family encoded by two δ subunits is much less well understood⁹. The existence of a large number of genes, coupled with alternative splicing¹⁰ and RNA editing^{11,12}, makes the functional analysis of native glutamate receptors a formidable task.

Biophysical studies have revealed many intriguing features for these channels, such as the unique requirement for coactivation of NMDA receptors by both glycine and glutamate, which bind to NR1 and NR2 subunits, respectively. But the most exciting recent advance is the solution of high-resolution crystal structures for the ligand-binding cores for individual iGluR subunits. This has been achieved for representative members from each gene family, providing for the first time a detailed understanding of agonist recognition in a diverse family of ligand-gated ion channels. From these structures, plausible models have been rigorously tested to identify in broad terms the mechanisms underlying activation and desensitization, and the action of some allosteric modulators on the ligand-binding domain.

In this review I discuss the molecular organization of iGluRs; I explain how crystallographic studies have shaped our thinking about how agonists open the channel; and I summarize what we know about the rapid desensitization and allosteric modulation of AMPA receptors.

Modular organization of iGluR subunits

The iGluRs have a characteristic modular architecture that seems to have been formed by fusion of three genetically discrete gene segments

that in bacteria encode individual proteins (Fig. 1). The amino-terminal domain (ATD) has homology with leucine-isoleucine-valine-binding protein (LIVBP), a bacterial periplasmic binding protein, as well as with the structurally related glutamate-binding domain of the G-protein-coupled metabotropic receptors. In contrast to its role in G-protein-coupled glutamate receptors, in iGluRs the ATD does not bind glutamate but instead forms dimers that are major assembly control points limiting oligomerization to members of the same iGluR subfamily^{13,14}. In NMDA receptors, the ATD of the NR2A and NR2B subunits bind the negative allosteric modulators Zn^{2+} and the drug ifenprodil, respectively^{15–17}. The results of site-directed mutagenesis and molecular modelling studies, which suggest that Zn^{2+} and ifenprodil promote domain closure and stabilize a closed cleft conformation of the ATD, raise the question

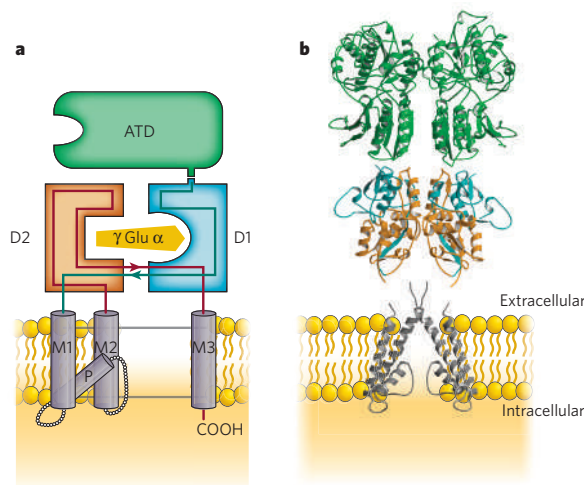


Figure 1 | Modular organization of a glutamate receptor. **a**, Diagrammatic representation of the amino-terminal domain (ATD), the two domain (D1 and D2) ligand-binding core encoded by the S1 (cyan) S2 (orange) peptide segments, the ion-channel domain with three membrane-spanning segments and a pore-loop (P), and the intracellular C terminus. **b**, Molecular architecture illustrated by the crystal structures of the mGluR1 ligand-binding domain dimer (green), the GluR5 ligand-binding core dimer (S1, cyan; S2 orange) and two subunits from the KcsA tetramer (grey). Note that S1 and S2 contribute to both domains of the ligand-binding core owing to the crossover of S1 into D2 and S2 into D1. An intact glutamate receptor is believed to assemble as a dimer of dimers.

¹Building 35, Room 3B1002, Porter Neuroscience Research Center, 35 Lincoln Drive, Bethesda, Maryland 20892, USA.

of whether there are as-yet-unidentified endogenous modulators that bind to other members of the iGluR family^{18,19}. It is also possible that the ATD of many iGluRs has lost the ability to bind endogenous ligands and that its major role now is to control assembly.

The agonist-binding domain of iGluRs is encoded by two polypeptide segments named S1 and S2, which as shown in Fig. 1 are interrupted by insertion of the ion-channel pore²⁰. This domain also has sequence and structural homology with bacterial periplasmic proteins but it differs in structure from LIVBP and the ATD of iGluRs. With genetic approaches it is possible to excise the agonist-binding domain from the remainder of an iGluR subunit and express just this as a soluble domain as a soluble protein suitable for crystallization^{21,22}. Many agonist and partial agonist structures have been solved, and this enormous body of new information is discussed in depth later in this review.

The ion-channel pore of iGluRs has striking sequence homology with that of bacterial potassium channels, suggesting that iGluRs also contain a pore-loop motif first identified in crystal structures of bacterial potassium channels^{23,24} (see also the review in this issue by Ashcroft, p. 440). This view was reinforced by the discovery of GluR0, a bacterial glutamate-gated potassium channel²⁵. However, in all iGluRs the channel orientation in the membrane is inverted compared with that found in classical potassium channels. Another important difference is that for all potassium channel structures solved so far, the subunits are arranged with perfect four-fold symmetry. In contrast, the agonist-binding domains of iGluRs are arranged as dimers of dimers. There is indirect evidence that for eukaryotic iGluRs, which do not discriminate between Na⁺ and K⁺, the extracellular pore entrance itself lacks four-fold symmetry²⁶.

The greatest diversity among iGluR subtypes is found in their cytoplasmic domain, which varies in size from less than 20 to around 500 amino acids. This is a key site of allosteric modulation by calmodulin, protein kinases and phosphatases^{27–29}. There is a large array of proteins that coassemble in the postsynaptic density, and many of these interact with glutamate receptors, either directly or through intermediate partners³⁰. Unfortunately very little is known about the molecular organization of this region. Algorithms for secondary-structure prediction indicate only short stretches of α -helix and β -sheet, suggesting that this region acts as more of a handle for cytoskeletal and cytoplasmic proteins, where the recognition sites are encoded not by classical three-dimensional structured domains, but rather by relatively short primary signature sequences. At both the structural and functional level this is a vast area to explore.

Ligand binding at atomic resolution

The general principle that glutamate, or glycine in the case of the NMDA receptor NR1 subunit, binds in the cleft of a clam-shell-like motif formed from the S1–S2 segments of a single iGluR subunit was anticipated from earlier structural analysis of bacterial periplasmic ligand-binding proteins. These served as the initial models for the ligand-binding domains of iGluRs^{31,32}. However, the chemistry of the ligand-binding pocket in periplasmic proteins, and even in the prokaryotic iGluR homologue GluR0 (ref. 33), differs appreciably from that for vertebrate iGluRs. What has emerged from crystallographic studies of the ligand-binding cores is atomic-resolution knowledge of the mechanisms that give each receptor its characteristic ligand-binding properties. This is an extraordinary advance that required the solution of multiple receptor–ligand complexes, the large majority at resolutions better than 2 Å, and at present the protein databank contains structures for more than 60 iGluR structures. Although the majority are for agonist and partial agonist complexes with GluR2, structures have also been solved for the kainate receptor GluR5 and GluR6 subunits, the NMDA receptor NR1 and NR2A subunits, and for several competitive antagonists.

In all of these structures an arginine side chain on helix D forms the major binding site for the ligand α -carboxyl group (Fig. 2). This residue is essential for the high-affinity binding of glutamate or glycine, and its mutation, even to lysine, greatly lowers agonist affinity. In all nine AMPA and kainate receptor genes, the ligand α -amino group is bound

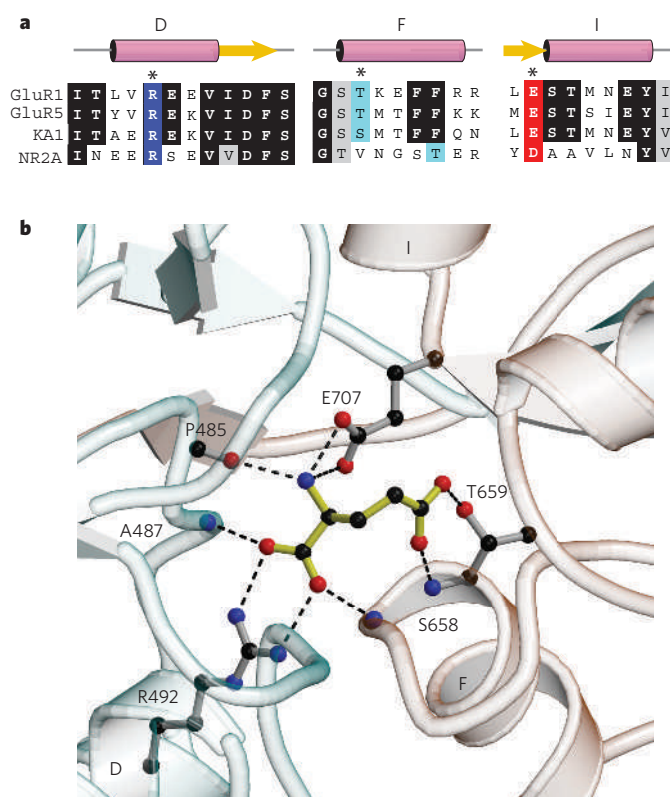


Figure 2 | Conserved structural elements in the agonist-binding site of iGluRs. **a**, The sequence alignment and secondary-structure assignment for GluR2 (α -helix in pink; β -sheet in yellow) indicate conserved features in the binding sites of AMPA, kainate and NMDA receptors. Key side chains that bind glutamate are highlighted in blue, red or cyan and marked with an asterisk; black and grey boxes indicate identical and conserved amino acids. **b**, The crystal structure for the GluR6 glutamate complex shows, as ball-and-stick models, the Arg (R492), Glu (E707) and Thr (T659) side chains that bind glutamate. Dashed lines indicate hydrogen bonds with main-chain and side-chain atoms. Not shown are four additional solvent-mediated hydrogen bonds that connect the agonist molecule with the ligand-binding site.

by a conserved glutamate side chain. In NMDA receptors, this glutamate is replaced by an aspartate, and in the NR2A complex the direct hydrogen bond to the agonist α -amino group is replaced by a solvent-mediated contact. Surprisingly, the γ -carboxyl group is not bound by a counter charge from a lysine or arginine side chain, but instead makes hydrogen-bond contacts with the main-chain peptide bond and the hydroxyl group of a conserved threonine side chain. The conformation adopted by glutamate is similar in AMPA, kainate and NMDA receptor NR2 subunits. As a result the γ -carboxyl group projects towards helix F in all iGluR structures. This finding was unexpected because structure–function studies with conformationally restricted ligands suggested that glutamate would adopt an extended conformation when bound to NMDA receptors but a folded conformation when bound to NMDA receptors³⁴.

Five key themes emerge from comparison of these structures. First, as shown in Fig. 3, the ligand-binding core is much larger than is necessary to accommodate glutamate. The exception is NR1, which binds glycine. NR1 has a binding pocket just about the right size for this smaller residue, and therefore excludes glutamate by steric effects³⁵. The cavity is completely closed off from the extracellular solution, except for NR2A, for which a solvent-filled tunnel connects the glutamate ligand in the interior of the protein to the extracellular environment. Second, subtype selectivity arises from amino-acid differences in domain 2, which shape the size and chemistry of the ligand-binding pocket; steric effects have key roles in permitting the binding of subtype-selective agonists

and excluding other ligands³⁶. In contrast, domain 1 shares a common shape and chemistry in all iGluR subtypes, consistent with its role in binding the α -amino and α -carboxyl groups, which are common to all iGluR agonists. Third, the extent of agonist-triggered domain closure varies between receptor subtypes, and also for individual agonists, partial agonists and antagonists^{36–39}. Fourth, competitive antagonists work by a foot-in-the-door mechanism, trapping the ligand-binding domain in a wide open conformation^{35,37,40}. Finally, the ligand-binding cavity contains trapped water molecules that have key roles in the agonist-binding mechanism.

Partial agonists

Partial agonists are ligands that when applied at a saturating concentration to a receptor produce less than the maximal response. Two complementary techniques have revealed how these ligands act. Single-channel recording can measure, on an absolute scale, the fraction of time that a channel spends in the open state⁴¹. In contrast, crystallographic structures reveal a snapshot of proteins caught in one of the many different conformations they sample during their functional life cycle. For AMPA receptors, these methods showed that agonist efficacy is directly coupled to the extent of domain closure³⁸. This conclusion was based on analysis of a series of 5-position halogen-substituted willedines, partial agonists for which the extent of domain closure varied owing to steric effects. Confirming this picture are additional crystal structures for individual structurally diverse partial agonists for both AMPA and kainate receptors including (S)-2-amino-3-(4-bromo-3-

hydroxy-isoxazol-5-yl)propionic acid and kainic acid^{36,39}. The appeal of this partial agonist mechanism is that it also explains the stepwise increase in single-channel conductance observed during dissociation of competitive antagonists from an AMPA receptor in the presence of a saturating concentration of glutamate⁴².

Mutation of Leu 650 to Tyr in the AMPA receptor reduced the steric hindrance that prevented kainate from acting as a full agonist. But this experiment also revealed a second mechanism closer in spirit to the widely accepted Monod–Wyman–Changeux (MWC) model for allosteric gating⁴³. The MWC-like mechanism, in which the affinity of ligands for a single active conformation determines their efficacy, was suggested by the observation that the L650T mutation reduced the efficacy but did not reduce the extent of domain closure for the complex with AMPA. This result hints that the reduced efficacy occurs because of a decreased stability of the fully activated AMPA-bound state. However, for kainate the L650T mutation increased both the efficacy and the extent of domain closure, indicating that ligands can activate receptors by more than one mechanism.

Subsequent studies on NMDA receptors have reinforced the view that different mechanisms can underlie partial agonist activity. The extent of domain closure for the NR1 subunit partial agonists ACPC and ACBC, which produce only 80% and 40% of the maximum response, is the same as that for full agonists⁴⁴. The high-resolution crystal structures solved for the NR1 subunit partial agonists revealed local conformational changes in the ligand-binding site, with the consequence that despite identical domain closure, the structures of the NR1 agonist and partial agonist complexes are indeed different. But the difference is much more subtle than observed for AMPA receptor partial agonists. Single-channel analysis of the partial agonist action of homoquinolinic acid on the glutamate-binding site of the NR2A subunit revealed a reduced rate constant for channel activation, which was linked to different binding mechanisms for glutamate and homoquinolinic acid, predicted by molecular dynamics (MD) simulations⁴⁵. Although the homology model developed for the NR2A glutamate complex differs in detail from the subsequently published NR2A crystal structure^{46,47}, and assumes that the extent of domain closure is identical for homoquinolinate and glutamate, the approach of combining high-resolution single-channel recording with MD simulations based on experimentally determined crystal structures has great potential. MD simulations on the AMPA-receptor glutamate complex, which measured the stability of hydrogen bonds and hydrophobic contacts, give a hint of what is possible for studies on the nanosecond timescale⁴⁸. More recent work, also on AMPA receptors, revealed for the first time transitions from the open to closed cleft conformation during 20-ns simulations⁴⁹. This work also revealed that as a result of differences in domain closure, water exchange rates within the ligand-binding pocket were much faster for the partial agonist kainate than for glutamate.

Coupling of ion-channel activity to agonist binding

A dimer is present in the vast majority of iGluR crystal structures, including the recently solved NMDA receptor NR1/NR2A heterodimer⁴⁷. As discussed below, such dimers have a key role in coupling the binding of glutamate to the activation of ion-channel gating. The dissociation constant, K_d , for dimer formation by the isolated ligand-binding cores is in the millimolar range for wild-type iGluRs, except for GluR0, for which the K_d is 0.8 μ M. Most probably, dimer formation occurs readily *in vitro* owing to the high protein concentrations present in crystals. This would be mimicked *in vivo* by the close proximity of the ligand-binding cores resulting from dimer formation by the ATDs that coassemble at much lower protein concentrations. In all subtypes of iGluR, the ligand-binding core dimers are 'glued' together by hydrophobic contacts made by the surface of helices D and J in domain 1, together with hydrogen bonds and salt bridges, the details of which vary according to receptor subtype. Notably, in the dimer assembly of AMPA and kainate receptors, domain 2 is free to move in the transition from the apo to agonist-bound conformation. Although the same is true in NMDA receptors, additional intersubunit contacts between

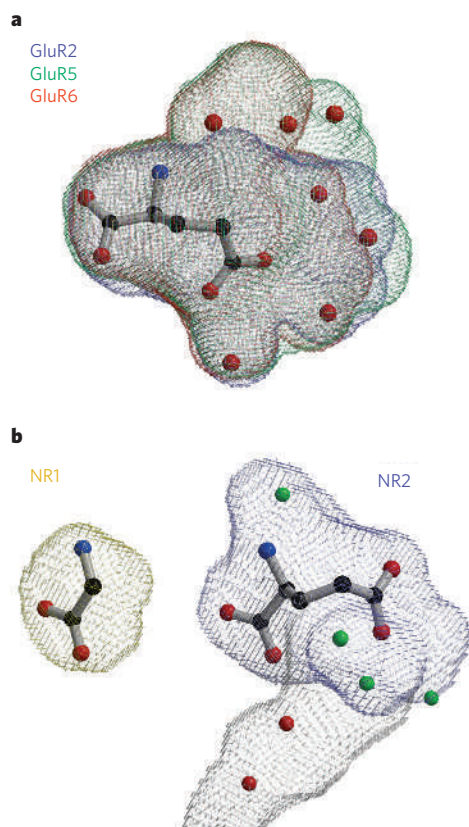


Figure 3 | Solvent-accessible surfaces of the ligand-binding cavity in iGluR agonist crystal structures. **a**, Superposition of the cavity interior in GluR2 (blue), GluR5 (green) and GluR6 (red); the bound ligand is glutamate plus seven water molecules shown as red spheres in the GluR5 complex. **b**, Cavities for NR1 (yellow) and NR2A (blue); the bound ligands are glycine for NR1 and glutamate plus four water molecules (green) in the NR2A complex; water molecules in the tunnel are shown as red spheres. Note that the glutamate ligand is much larger than can be accommodated by the NR1 cavity.

domains 1 and 2 suggest mechanisms for allosteric coupling between the glutamate- and glycine-binding sites.

The structure of the dimer assembly in iGluRs immediately suggested a mechanism for ion-channel gating. In this model, the movement of domain 2 towards domain 1 during the transition from the open cleft of the apo state to the closed cleft of the glutamate-bound complex produces a scissors-like outward motion of the linkers connecting the ion-channel transmembrane (TM) segments to the ligand-binding core (Fig. 4). Competitive antagonists jam the ligand-binding cores in their resting conformation by a foot-in-the-door mechanism. Although many details remain to be resolved, it seems reasonable to assume that gating of the ion channel in iGluRs results from movement of the TM segments that are pulled open by contraction of the ligand-binding cores in a dimer assembly. Because each subunit contains a complete agonist-binding site, which is coupled to TM segments in the same subunit, it is easy to imagine how the activation of individual subunits could generate partial opening of the ion channel. This would lead to the generation of multiple subconductance states characteristic of AMPA and kainate subtype iGluRs. This mechanism is similar to that subsequently proposed for the Ca^{2+} -activated K channel MthK^{33,37,50}. Although this principle is probably the same for NMDA receptors the details are much less well understood. Binding of either glutamate or glycine alone does not trigger ion-channel gating, and when the channel does open, it visits only one subconductance state, suggesting a cooperative gating scheme involving simultaneous, or tightly coupled, transitions by pairs of channel gates^{51,52}.

Desensitization and allosteric modulation

The gating model shown in Fig. 4 can explain how the channel opens because dimer assembly occurs mainly through the domain 1 surface, leaving domain 2 free to move. This suggests a model in which desensitization occurs when the dimer undergoes a conformational rearrangement that breaks contacts at the domain 1 surface. This uncoupling allows the pair of subunits in a dimer to rotate around the central axis, relieving strain on the ion-channel linkers and allowing the ion channel to close, even though the agonist remains bound with high affinity. This model was tested by constructing mutants that either increased or decreased the stability of dimer formation. The effects on dimer stability measured for isolated S1–S2 constructs correlated with the extent of desensitization measured in intact receptors bearing the same mutants⁵³. A subsequent systematic analysis of dimer contacts observed in the crystal structure was performed by site-directed mutagenesis, and changes in the kinetics of desensitization were measured⁵⁴. These results confirmed that in AMPA receptors the dimer interface must undergo a rearrangement in order for desensitization to proceed. It also verified that the stability of the dimer assembly in the resting and active state is generated by a network of intermolecular contacts that are distributed over an area of about 900–1400 Å². A major challenge for the future will be to identify the conformation of the ligand-binding cores in the desensitized state of a dimer assembly. An intriguing extension of this model, which could explain negative allosteric modulation by ligands such as Zn^{2+} and ifenprodil, is possible if one considers that domain closure of the ATD will pull apart the protomers in a ligand-binding core dimer assembly (Fig. 4).

The profound desensitization of AMPA receptors is strongly attenuated by small-molecule, positive allosteric modulators that seem to act through two somewhat related mechanisms. At the macroscopic level this can be modelled as resulting from either block of entry into the desensitized state (as occurs for cyclothiazide) or stabilization of the open state (as occurs for aniracetam⁵⁵). Biochemical and crystallographic experiments have established that both classes of compound bind in a solvent-filled hydrophobic crevice at the base of the ligand-binding core dimer, but at different sites. In sedimentation equilibrium experiments using purified GluR2 ligand-binding cores, both cyclothiazide and CX614 greatly stabilized the dimer. Cyclothiazide binds to a pair of sites at the periphery of the dimer, whereas aniracetam and CX614 bind to a single site located on the molecular two-fold axis of

symmetry^{53,56}. Cyclothiazide slows desensitization by greatly stabilizing the dimer interface, whereas aniracetam slows deactivation by stabilizing the ligand-binding domain in the agonist-bound, closed-cleft, active conformation. The above studies are striking examples of how the combination of crystallographic and functional studies can dissect complex allosteric mechanisms. Indeed, iGluRs are the only ligand-gated ion channels for which there is any structural basis for the mechanism of action of allosteric modulators.

The process of gating and desensitization is of course more complex than can be explained by the simple two-state models shown in Fig. 4. Mutations outside the dimer interface, in the linker from S2 to the second transmembrane segment, attenuate desensitization in both AMPA and kainate receptors. But in the absence of structural data, it is not possible to explain how this occurs⁵⁷. Mutations in the ligand-binding pocket can also profoundly affect desensitization: notably, replacement by alanine of a tyrosine side chain that caps the binding site in AMPA receptors converts kainate from a very weak partial agonist into a potent and strongly desensitizing ligand, most probably because of relief of steric hindrance and increased domain closure⁵⁸. Here the mechanism is easier to conceptualize, although the lack of structural data for the alanine mutant means that a definitive interpretation is not possible. The same mutation also greatly reduces the potency for glutamate, possibly as a result of disruption of hydrogen-bond contacts and solvent structure in the ligand-binding site.

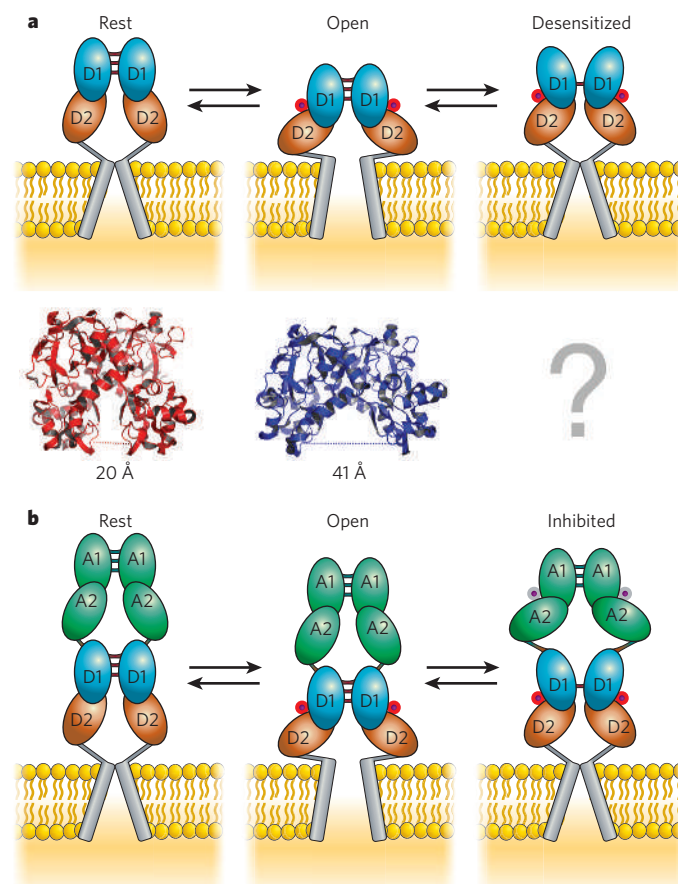


Figure 4 | Activation of iGluR gating, resulting from agonist-induced expansion of the ligand-binding core dimer. **a**, Diagrammatic representations of the resting, activated and desensitized states to illustrate how domain closure is linked to separation of the ligand-binding core dimer domain 2 segments. Lower panel, crystal structures of the GluR5 ligand-binding core dimer in the resting (antagonist bound) and active glutamate-bound conformations⁴⁰. Currently no structure for the desensitized state has been reported. **b**, Diagrammatic representations indicating a hypothetical model of how binding of an allosteric modulator to the ATD could pull the agonist-binding core dimer apart and trigger channel closure. Small red circles represent glutamate.

NMDA-receptor dimer assemblies

The recent crystallization of a heterodimer of the ligand-binding cores of the NR1–NR2A complex, and the biochemical demonstration using disulphide-linked receptors that this assembly is the native form, as opposed to separate NR1 and NR2 homodimers, is a major achievement that explains several unique functional properties of NMDA receptors⁴⁷. Although functional experiments have yet to begin in earnest, two results are of particular note. First, the slow deactivation of NMDA receptors, which is key to their biological function at synapses⁵⁹, occurs in part as a result of an amino-acid difference in the NR1 subunit in which a tyrosine residue at the base of the dimer interface occupies the site to which aniracetam binds in AMPA receptors. The tyrosine thus acts as an endogenous allosteric modulator that slows deactivation⁴⁷. Second, in contrast to the situation for AMPA and kainate receptors, in which the ligand-binding core dimer interface is generated exclusively by domain 1 contacts, in NMDA receptors both domain 1 and domain 2 contribute to the dimer interface. This suggests a mechanism for allosteric coupling between NR2 and NR1, which underlies the large variation in glycine affinity observed on coassembly with the NR2A–NR2D subunits. Experimental tests to define exactly how this occurs have yet to be reported^{60,61}.

The ion channel

A variety of approaches subsequently revealed the transmembrane topology of iGluRs. From a structural perspective the most fruitful have been cysteine-scanning mutagenesis experiments in which the rate of covalent modification of cysteine side chains was measured^{62,63}. This information was used to infer the structure of the channel. These experiments have reached a high degree of sophistication such that the voltage dependence of modification has been measured to locate the position of side chains in the membrane electric field⁶⁴. In conceptually related experiments, scanning mutagenesis with glutamate, alanine and tryptophan residues has given clues about the surface topology at the

cytoplasmic entrance to the pore and the location of the binding site for polyamines, which act as endogenous pore blockers for AMPA and kainate receptors with a glutamine at the Q/R editing site²³.

The picture that has emerged from these studies (Fig. 5) suggests a three-helix bundle supporting a pore-loop motif that forms the closest point of contact with ions near the cytoplasmic entrance to the channel⁶⁵. The ion-channel pore is lined mainly by residues from M2, the second transmembrane segment, with Ca^{2+} permeability determined both by the amino acid at the tip of the pore loop (the Q/R site) and by the DRPEER sequence in the linker between M2 and S2 in NMDA receptors⁶⁶. Until recently little attention was given to the M3 segment, which is present in all eukaryotic iGluRs but not their prokaryotic homologues. In addition, its location relative to M1 and M2 is uncertain. But M3 is vital for channel activity in eukaryotes, and its removal, together with the associated cytoplasmic C terminus, produces inactive receptors. Surprisingly, coexpression of M3 and truncated receptors restores activity, showing that there is no need for a covalent link between M3 and S2 during receptor activation⁶⁷. Also significant is the observation that the GYKI family of noncompetitive AMPA receptor antagonists seem to interact with the S2–M3 linker⁶⁸.

The gating model for iGluRs, based on movements of the linker regions in the ligand-binding core crystal dimers, suggests that the top of the channel, near the extracellular surface, must undergo large conformational movements. Unfortunately the details are fuzzy and even the location of the barrier to ion permeation in the closed and desensitized states remains to be unambiguously determined. The dimer of dimers model implied by the crystal structures of the ligand-binding cores has not been confirmed at a structural level, and no tetrameric assembly has yet been visualized at a resolution sufficient to resolve how the individual subunits are positioned with respect to each other. Despite this, evidence for a dimer of dimers is now overwhelming. This potentially forces the issue of a mismatch between the expected four-fold symmetry of the ion-channel pore, and the two-fold symmetry required by the dimer of dimers arrangement, but only if the ion-channel pore has four-fold symmetry. State-dependent crosslinking experiments on AMPA receptors suggest that the top of the helix-bundle crossing implied by a KcsA-based structural model is better modelled as a dimer of dimers rather than a tetramer, indicating that if a symmetry mismatch does occur its location is deeper in the channel²⁶. However, for NMDA receptors in which the NR1 and NR2 subunits form obligate heteromeric assemblies, the symmetry mismatch goes as far as the tip of the pore loop. This is not surprising given the different amino-acid sequence in the NR1 and NR2 subunits in this region. In retrospect there is no compelling reason to expect four-fold symmetry within the ion-channel pore for any eukaryotic glutamate receptor. However, for GluR0, which is highly potassium-selective, the situation is less clear. All of the K-channel structures solved so far exhibit four-fold symmetry, but coordination of K^+ ions does not require this *per se*, so perhaps the structure of even the GluR0 pore is somewhat different.

Glutamate receptors and disease

The availability of selective and potent NMDA receptor antagonists, especially the channel blocker MK801, triggered an enormous research effort in the 1980s when it was discovered that excessive calcium influx through NMDA receptors triggered neurodegeneration in animal models of ischaemia. The initial surge of excitement was rapidly tempered by studies that showed that NMDA receptor antagonists had severe behavioural side effects. This, coupled with increasingly sophisticated research on stroke mechanisms that revealed a pathophysiology much more complicated than could be explained by hyperactivation of NMDA receptors alone, greatly diminished enthusiasm for further work with glutamate-receptor antagonists as neuroprotective agents. However, the issue is not dead, and there is reasonable hope that, by fine-tuning receptor-subtype selectivity and affinity, clinically useful drugs will eventually emerge. Impetus for this is a growing body of evidence that glutamate receptors have key roles in a much broader variety of neurodegenerative and psychiatric diseases than first anticipated and

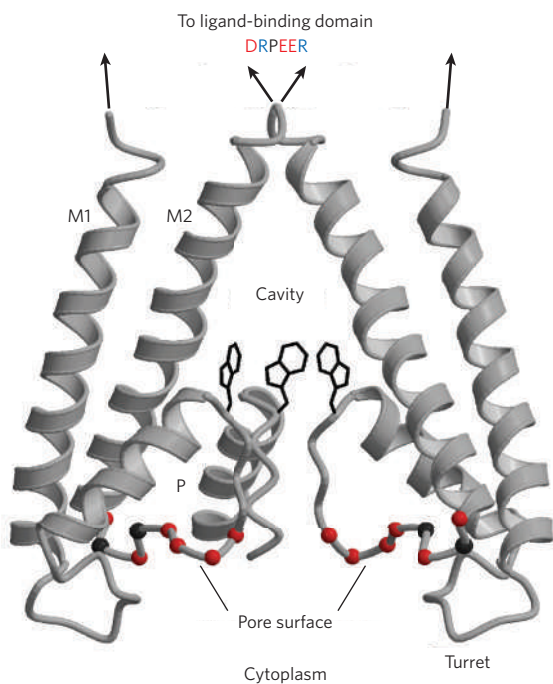


Figure 5 | The pore region of the KcsA potassium channel used as a template to map the results of site-directed mutagenesis on iGluR channel block by polyamines. At the tip of the pore loop (P), but nowhere else in the channel, introduction of tryptophan produces a pM affinity-binding site for polyamines²³. At the cytoplasmic entrance to the channel, red spheres indicate amino-acid α -carbon atom positions for which introduction of glutamate side chains supports polyamine block. In the linker from M2 to the ligand-binding domain, the DRPEER sequence probably contributes to a Ca^{2+} -binding site near the entrance to the channel⁶⁶.

the approval of the NMDA-receptor channel blocker Memantine for treatment of dementia in the United States and the European Union⁶⁹.

Surprisingly, in humans there are no diseases that have been unambiguously linked to mutations in glutamate-receptor genes, although several animal diseases have been discovered, which have provided us with useful mouse models. Two of the most interesting are *lurcher* and *stargazer*. *Lurcher* is produced by a point mutation in the second TM segment of the $\delta 2$ subunit. This mutation causes spontaneous ion-channel activity, possibly as a result of attenuated desensitization, which leads to excessive cation flux and neurodegeneration of Purkinje cells, which express $\delta 2$ at high levels^{70–72}. The *stargazer* phenotype results from a loss of AMPA-receptor trafficking to synapses on cerebellar granule cells, which has been traced to a mutation in the $\gamma 2$ member of a family of auxiliary membrane proteins, which modulate both AMPA receptor trafficking and gating kinetics^{73–75}. The finding that AMPA receptors *in vivo* coassemble with an auxiliary subunit, and that this modifies their affinity for agonists and their kinetics of deactivation and desensitization, was entirely unexpected.

Future prospects

The extraordinary advances in understanding the ligand-binding mechanisms of iGluRs arose largely from crystallographic efforts. Although many important details will be learned in the immediate future from crystallization of the ligand-binding cores of additional iGluR subtypes and complexes, the major challenge is to extend this approach to other receptor domains, perhaps to an intact receptor, and to figure out how the domains communicate with each other. As a consequence of our limited ability to overexpress and purify eukaryotic membrane proteins, it is far from clear when or if this will be possible. As a result, details of the ion-permeation process, the binding sites for channel blockers, and the site of interactions with recently identified accessory membrane proteins remain to be determined. In this sense, our understanding of Cys-loop receptors, for which structures are available for the intact protein, albeit at lower resolution, is more advanced (see the review by Sine and Engel, p. 448 in this issue). Even with (or without) this information, studies using a variety of experimental approaches including calorimetry, fluorescence resonance energy transfer (FRET), electron microscopy, NMR, and of course electrophysiological techniques, will be required to bring life to the static iGluR structures revealed by crystallography. Further investment in computational studies is another approach with great potential, but major advances will require the development of techniques to describe protein motion on the microsecond-to-millisecond time scale. The past five years have been incredibly productive and I hope progress will continue at this rate as more complex structural problems are tackled. ■

- Watkins, J. C. & Evans, R. H. Excitatory amino acid transmitters. *Annu. Rev. Pharmacol. Toxicol.* **21**, 165–204 (1981).
- Mayer, M. L. & Westbrook, G. L. The physiology of excitatory amino acids in the vertebrate central nervous system. *Prog. Neurobiol.* **28**, 197–276 (1987).
- Collingridge, G. L. & Lester, R. A. J. Excitatory amino acid receptors in the vertebrate central nervous system. *Pharmacol. Rev.* **41**, 143–210 (1989).
- Collingridge, G. L. & Bliss, T. V. Memories of NMDA receptors and LTP. *Trends Neurosci.* **18**, 54–56 (1995).
- Geiger, J. R. P. *et al.* Relative abundance of subunit mRNAs determines gating and Ca^{2+} permeability of AMPA receptors in principle neurons and interneurons in rat CNS. *Neuron* **15**, 193–204 (1995).
- Lerma, J. Roles and rules of kainate receptors in synaptic transmission. *Nature Rev. Neurosci.* **4**, 481–495 (2003).
- Perl, T. M. *et al.* An outbreak of toxic encephalopathy caused by eating mussels contaminated with domoic acid. *N. Engl. J. Med.* **322**, 1775–1780 (1990).
- Watkins, J. C. & Jane, D. E. The glutamate story. *Br. J. Pharmacol.* **147** Suppl 1, S100–S108 (2006).
- Hollmann, M. in *Ionotropic Glutamate Receptors in the CNS* (eds Jonas, P. & Monyer, H.) 3–98 (Springer, Berlin, 1999).
- Sommer, B. *et al.* Flip and flop: a cell-specific functional switch in glutamate-operated channels of the CNS. *Science* **249**, 1580–1585 (1990).
- Sommer, B., Köhler, M., Sprengel, R. & Seeburg, P. H. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67**, 11–19 (1991).
- Puchalski, R. B. *et al.* Selective RNA editing and subunit assembly of native glutamate receptors. *Neuron* **13**, 131–147 (1994).
- Ayalon, G. & Stern-Bach, Y. Functional assembly of AMPA and kainate receptors is mediated by several discrete protein-protein interactions. *Neuron* **31**, 103–113 (2001).
- Ayalon, G., Segev, E., Elgavish, S. & Stern-Bach, Y. Two regions in the N-terminal domain of ionotropic glutamate receptor 3 form the subunit oligomerization interfaces that control subtype-specific receptor assembly. *J. Biol. Chem.* **280**, 15053–15060 (2005).
- Paoletti, P., Ascher, P. & Neyton, J. High-affinity zinc inhibition of NMDA NR1-NR2A receptors. *J. Neurosci.* **17**, 5711–5725 (1997).
- Perin-Dureau, F., Rachline, J., Neyton, J. & Paoletti, P. Mapping the binding site of the neuroprotectant ifenprodil on NMDA receptors. *J. Neurosci.* **22**, 5955–5965 (2002).
- Masuko, T. *et al.* A regulatory domain (R1-R2) in the amino terminus of the N-methyl-D-aspartate receptor: effects of spermine, protons, and ifenprodil, and structural similarity to bacterial leucine/isoleucine/valine binding protein. *Mol. Pharmacol.* **55**, 957–969 (1999).
- Paoletti, P. *et al.* Molecular organization of a zinc binding N-terminal modulatory domain in a NMDA receptor subunit. *Neuron* **28**, 911–925 (2000).
- Pasternack, A. *et al.* Alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor channels lacking the N-terminal domain. *J. Biol. Chem.* **277**, 49662–49667 (2002).
- Stern-Bach, Y. *et al.* Agonist-selectivity of glutamate receptors is specified by two domains structurally related to bacterial amino acid binding proteins. *Neuron* **13**, 1345–1357 (1994).
- Kuusinen, A., Arvola, M. & Keinänen, K. Molecular dissection of the agonist binding site of an AMPA receptor. *EMBO J.* **14**, 6327–6332 (1995).
- Chen, G. Q., Sun, Y., Jin, R. & Gouaux, E. Probing the ligand binding domain of the GluR2 receptor by proteolysis and deletion mutagenesis defines domain boundaries and yields a crystallizable construct. *Protein Sci.* **7**, 2623–2630 (1998).
- Panchenko, V. A., Glasser, C. R. & Mayer, M. L. Structural similarities between glutamate receptor channels and K^+ channels examined by scanning mutagenesis. *J. Gen. Physiol.* **117**, 345–360 (2001).
- Kuner, T., Seeburg, P. H. & Guy, H. R. A common architecture for K^+ channels and ionotropic glutamate receptors? *Trends Neurosci.* **26**, 27–32 (2003).
- Chen, G. Q., Cui, C., Mayer, M. L. & Gouaux, E. Functional characterization of a potassium-selective prokaryotic glutamate receptor. *Nature* **402**, 817–821 (1999).
- Sobolevsky, A. I., Yelshansky, M. V. & Wollmuth, L. P. The outer pore of the glutamate receptor channel has 2-fold rotational symmetry. *Neuron* **41**, 367–378 (2004).
- Soderling, T. R. & Derkach, V. A. Postsynaptic protein phosphorylation and LTP. *Trends Neurosci.* **23**, 75–80 (2000).
- Tavalin, S. J. *et al.* Regulation of GluR1 by the A-kinase anchoring protein 79 (AKAP79) signaling complex shares properties with long-term depression. *J. Neurosci.* **22**, 3044–3051 (2002).
- Vissel, B., Krupp, J. J., Heinemann, S. F. & Westbrook, G. L. Intracellular domains of NR2 alter calcium-dependent inactivation of N-methyl-D-aspartate receptors. *Mol. Pharmacol.* **61**, 595–605 (2002).
- Peng, J. *et al.* Semiquantitative proteomic analysis of rat forebrain postsynaptic density fractions by mass spectrometry. *J. Biol. Chem.* **279**, 21003–21011 (2004).
- Mano, I., Lamed, Y. & Teichberg, V. I. A venus flytrap mechanism for activation and desensitization of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors. *J. Biol. Chem.* **271**, 15299–15302 (1996).
- Quirocho, F. A. & Ledvina, P. S. Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: variation of common themes. *Mol. Microbiol.* **20**, 17–25 (1996).
- Mayer, M. L., Olson, R. & Gouaux, E. Mechanisms for ligand binding to GluR0 ion channels: crystal structures of the glutamate and serine complexes and a closed apo state. *J. Mol. Biol.* **311**, 815–836 (2001).
- Watkins, J. C., Krogsgaard-Larsen, P. & Honoré, T. Structure-activity relationships in the development of excitatory amino acid receptor agonists and competitive antagonists. *Trends Pharmacol. Sci.* **11**, 25–33 (1990).
- Furukawa, H. & Gouaux, E. Mechanisms of activation, inhibition and specificity: crystal structures of NR1 ligand-binding core. *EMBO J.* **22**, 1–13 (2003).
- Mayer, M. L. Crystal structures of the GluR5 and GluR6 ligand binding cores: molecular mechanisms underlying kainate receptor selectivity. *Neuron* **45**, 539–552 (2005).
- Armstrong, N. & Gouaux, E. Mechanisms for activation and antagonism of an AMPA-sensitive glutamate receptor: Crystal structures of the GluR2 ligand binding core. *Neuron* **28**, 165–181 (2000).
- Jin, R., Banke, T. G., Mayer, M. L., Traynelis, S. F. & Gouaux, E. Structural basis for partial agonist action at ionotropic glutamate receptors. *Nature Neurosci.* **6**, 803–810 (2003).
- Hogner, A. *et al.* Structural basis for AMPA receptor activation and ligand selectivity: Crystal structures of five agonist complexes with the GluR2 ligand-binding core. *J. Mol. Biol.* **322**, 93–109 (2002).
- Mayer, M. L., Ghosal, A. K., Dolman, N. P. & Jane, D. E. Crystal structures of the kainate receptor GluR5 ligand binding core dimer with novel GluR5 selective antagonists. *J. Neurosci.* **26**, 2850–2859 (2006).
- Colquhoun, D. & Ogden, D. C. Activation of ion channels in the frog end-plate by high concentrations of acetylcholine. *J. Physiol. (Lond.)* **395**, 31–159 (1988).
- Rosenmund, C., Stern-Bach, Y. & Stevens, C. F. The tetrameric structure of a glutamate receptor channel. *Science* **280**, 1596–1599 (1998).
- Armstrong, N., Mayer, M. & Gouaux, E. Tuning activation of the AMPA-sensitive GluR2 ion channel by genetic adjustment of agonist-induced conformational changes. *Proc. Natl Acad. Sci. USA* **100**, 5736–5741 (2003).
- Inanobe, A., Furukawa, H. & Gouaux, E. Mechanism of partial agonist action at the NR1 subunit of NMDA receptors. *Neuron* **47**, 71–84 (2005).
- Erreger, K. *et al.* Mechanism of partial agonism at NMDA receptors for a conformationally restricted glutamate analog. *J. Neurosci.* **25**, 7858–7866 (2005).
- Chen, P. E. *et al.* Structural features of the glutamate binding site in recombinant NR1/NR2A N-methyl-D-aspartate receptors determined by site-directed mutagenesis and molecular modeling. *Mol. Pharmacol.* **67**, 1470–1484 (2005).
- Furukawa, H., Singh, S. K., Mancusso, R. & Gouaux, E. Subunit arrangement and function in NMDA receptors. *Nature* **438**, 185–192 (2005).
- Mamonova, T., Hespeneheide, B., Straub, R., Thorpe, M. F. & Kurnikova, M. Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.* **2**, S137–S147 (2005).
- Arinaminpathy, Y., Sansom, M. S. & Biggin, P. C. Binding site flexibility: Molecular simulation of partial and full agonists with a glutamate receptor. *Mol. Pharmacol.* **69**, 5–12 (2006).

50. Jiang, Y. *et al.* The open pore conformation of potassium channels. *Nature* **417**, 523–526 (2002).
51. Banke, T. G. & Traynelis, S. F. Activation of NR1/NR2B NMDA receptors. *Nature Neurosci.* **6**, 144–152 (2003).
52. Schorge, S., Elenes, S. & Colquhoun, D. Maximum likelihood fitting of single channel NMDA activity with a mechanism composed of independent dimers of subunits. *J. Physiol.* **569**, 395–418 (2005).
53. Sun, Y. *et al.* Mechanism of glutamate receptor desensitization. *Nature* **417**, 245–253 (2002).
54. Horning, M. S. & Mayer, M. L. Regulation of AMPA receptor gating by ligand binding core dimers. *Neuron* **41**, 379–388 (2004).
55. Partin, K. M., Fleck, M. W. & Mayer, M. L. AMPA receptor flip/flop mutants affecting deactivation, desensitization, and modulation by cyclothiazide, aniracetam, and thiocyanate. *J. Neurosci.* **16**, 6634–6647 (1996).
56. Jin, R. *et al.* Mechanism of positive allosteric modulators acting on AMPA receptors. *J. Neurosci.* **25**, 9027–9036 (2005).
57. Yelshansky, M. V., Sobolevsky, A. I., Jatzke, C. & Wollmuth, L. P. Block of AMPA receptor desensitization by a point mutation outside the ligand-binding domain. *J. Neurosci.* **24**, 4728–4736 (2004).
58. Holm, M. M. *et al.* A binding site tyrosine shapes desensitization kinetics and agonist potency at GluR2. A mutagenic, kinetic, and crystallographic study. *J. Biol. Chem.* **280**, 35469–35476 (2005).
59. Lester, R. A. J., Clements, J. D., Westbrook, G. L. & Jahr, C. E. Channel kinetics determine the time course of NMDA receptor-mediated synaptic currents. *Nature* **346**, 565–567 (1990).
60. Monyer, H. *et al.* Heteromeric NMDA receptors: molecular and functional distinction of subtypes. *Science* **256**, 1217–1221 (1992).
61. Laurie, D. J. & Seeburg, P. H. Ligand affinities at recombinant N-methyl-D-aspartate receptors depend on subunit composition. *Eur. J. Pharmacol.* **268**, 335–345 (1994).
62. Kuner, T., Wollmuth, L. P., Karlin, A., Seeburg, P. H. & Sakmann, B. Structure of the NMDA receptor channel M2 segment inferred from the accessibility of substituted cysteines. *Neuron* **17**, 343–352 (1996).
63. Beck, C., Wollmuth, L. P., Seeburg, P. H., Sakmann, B. & Kuner, T. NMDAR channel segments forming the extracellular vestibule inferred from the accessibility of substituted cysteines. *Neuron* **22**, 559–570 (1999).
64. Sobolevsky, A. I., Yelshansky, M. V. & Wollmuth, L. P. State-dependent changes in the electrostatic potential in the pore of a GluR channel. *Biophys. J.* **88**, 235–242 (2005).
65. Wollmuth, L. P. & Sobolevsky, A. I. Structure and gating of the glutamate receptor ion channel. *Trends Neurosci.* **27**, 321–328 (2004).
66. Watanabe, J., Beck, C., Kuner, T., Premkumar, L. S. & Wollmuth, L. P. DRPEER: a motif in the extracellular vestibule conferring high Ca^{2+} flux rates in NMDA receptor channels. *J. Neurosci.* **22**, 10209–10216 (2002).
67. Schorge, S. & Colquhoun, D. Studies of NMDA receptor function and stoichiometry with truncated and tandem subunits. *J. Neurosci.* **23**, 1151–1158 (2003).
68. Balannik, V., Menniti, F. S., Paternain, A. V., Lerma, J. & Stern-Bach, Y. Molecular mechanism of AMPA receptor noncompetitive antagonism. *Neuron* **48**, 279–288 (2005).
69. Lipton, S. A. Paradigm shift in neuroprotection by NMDA receptor blockade: Memantine and beyond. *Nature Rev. Drug Discov.* **5**, 160–170 (2006).
70. Zuo, J. *et al.* Neurodegeneration in Lurcher mice caused by mutation in delta2 glutamate receptor gene. *Nature* **388**, 769–773 (1997).
71. Kohda, K., Wang, Y. & Yuzaki, M. Mutation of a glutamate receptor motif reveals its role in gating and delta2 receptor channel properties. *Nature Neurosci.* **3**, 315–322 (2000).
72. Klein, R. M. & Howe, J. R. Effects of the lurcher mutation on GluR1 desensitization and activation kinetics. *J. Neurosci.* **24**, 4941–4951 (2004).
73. Tomita, S. *et al.* Stargazin modulates AMPA receptor gating and trafficking by distinct domains. *Nature* **435**, 1052–1058 (2005).
74. Turetsky, D., Garringer, E. & Patneau, D. K. Stargazin modulates native AMPA receptor functional properties by two distinct mechanisms. *J. Neurosci.* **25**, 7438–7448 (2005).
75. Priel, A. *et al.* Stargazin reduces desensitization and slows deactivation of the AMPA-type glutamate receptors. *J. Neurosci.* **25**, 2682–2686 (2005).

Acknowledgements Work in the author's laboratory is supported by the intramural research programme of NICHD, NIH, DHHS. Synchrotron diffraction data were collected at Southeast Regional Collaborative Access Team (SER-CAT) 22-ID beamline at the Advanced Photon Source, Argonne National Laboratory. Use of the Advanced Photon Source was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences.

Author Information Reprints and permissions information is available at npg.nature.com/reprints and permissions. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to the author (mayerm@mail.nih.gov).

hERG potassium channels and cardiac arrhythmia

Michael C. Sanguinetti^{1,2} & Martin Tristani-Firouzi^{2,3}

hERG potassium channels are essential for normal electrical activity in the heart. Inherited mutations in the *HERG* gene cause long QT syndrome, a disorder that predisposes individuals to life-threatening arrhythmias. Arrhythmia can also be induced by a blockage of hERG channels by a surprisingly diverse group of drugs. This side effect is a common reason for drug failure in preclinical safety trials. Insights gained from the crystal structures of other potassium channels have helped our understanding of the block of hERG channels and the mechanisms of gating.

Potassium-selective channels have diverse structures and functions. Roles for them include maintenance of the resting membrane potential in all cell types and termination of action potentials in excitable cells. The importance of hERG (human *ether-a-go-go*-related gene¹) K⁺ channels in normal human cardiac electrical activity became strikingly obvious when inherited mutations in *HERG* were found to cause long QT syndrome (LQTS)², a cardiac repolarization disorder that predisposes affected individuals to arrhythmia (rapid irregular heart beats that can lead to fainting and sudden death). Before this discovery, it was known that a similar disorder could be induced by high doses of some hERG-channel blockers (for example, quinidine and dofetilide) that were, ironically, used to prevent arrhythmia. Other common non-cardiac medications (such as some antihistamines and antibiotics) can also trigger ventricular arrhythmia and sudden death³ via block of the hERG channel. This led to their withdrawal by drug regulatory agencies. It is now routine practice in the pharmaceutical industry to test compounds for hERG-channel activity early in the drug-development process.

In this review we summarize the normal physiological functions of hERG channels, define the structural basis of their unusual gating and describe how mutations in the *HERG* gene or promiscuous blocking of the channels by drugs cause arrhythmia. Much of our current understanding of hERG-channel gating is inferred from detailed biophysical studies of another voltage-gated K⁺ (Kv) channel, called Shaker, in *Drosophila*. In addition, recent X-ray crystal structures of several bacterial K⁺ channels^{4–6} and a mammalian Kv1.2 K⁺ channel⁷ have revolutionized our understanding of the structural basis of selective K⁺ permeability and channel gating. In the absence of a crystal structure for hERG, analysis of mutant channels has provided useful insights into their function.

hERG mediates cardiac repolarization

hERG is expressed in multiple tissues and cell types, including neural¹, smooth muscle⁸ and tumour cells^{9,10}. In the central nervous system of mammals, two other ERG channel genes are expressed (*ERG2* and *ERG3*)¹¹, and here heteromultimeric channel complexes of all three subunits can be formed¹². However, hERG is most highly expressed in the heart, and this is where its function — and dysfunction — is best understood.

The action potential of human ventricular myocytes can be divided into five distinct phases (phases 0 to 4; Fig. 1a). Activation of inward Na⁺ current triggers a rapid depolarization of the membrane (phase 0).

Repolarization is much slower and occurs in three phases. Phase 1 repolarization proceeds rapidly, lasts only a few milliseconds and is followed by a much slower rate of repolarization (phase 2) called the plateau. The plateau of cardiac action potentials is prolonged because the K⁺ currents activated during this phase are slow to activate and/or have a reduced conductance at positive transmembrane potentials (Fig. 1b). A long action potential ensures adequate time for entry of extracellular Ca²⁺ into the myocyte for optimum excitation–contraction coupling. Delayed repolarization also makes cardiac muscle refractory to premature excitation — an important, but imperfect, safeguard against the generation of re-entrant arrhythmia. The third phase of repolarization terminates the action potential and returns the membrane potential to its resting level (phase 4). The most important component of phase 3 repolarization is the rapid delayed rectifier K⁺ current (*I_{Kr}*) conducted by hERG channels^{13,14}.

At negative membrane potentials (for example, –80 mV), hERG channels are in a non-conducting closed state (Fig. 2). Depolarization of the cell membrane to potentials more positive than –60 mV induces channels to open and allows the outward diffusion of K⁺ ions in accordance with its electrochemical driving force. As the membrane potential is progressively depolarized to more positive potentials, channels enter another non-conducting configuration called the inactivated state. hERG gating is unlike that of most other Kv channels in two ways. First, inactivation is much faster than activation: this reduces outward conductance at depolarized potentials and prolongs phase 2 of the action potential. Second, recovery from inactivation is much faster than deactivation: this initiates phase 3 repolarization of the action potential (Fig. 1a).

Mutations in hERG cause cardiac arrhythmia

LQTS is defined by prolongation of the QT interval measured by a body surface electrocardiogram (ECG) and a greatly increased risk of ventricular fibrillation. The QT interval is the time required for ventricular repolarization during a single cardiac cycle (Fig. 3a, b). Delayed repolarization increases the risk of *torsade de pointes* (TdP), a unique cardiac arrhythmia characterized by an ECG trace that resembles a waxing and waning sine wave (Fig. 3c). TdP can either revert to a normal sinus rhythm or degenerate into lethal ventricular fibrillation. LQTS affects an estimated 1 in 5,000–10,000 people worldwide and is caused most often by dominant mutations in *HERG* or *KCNQ1*, the α -subunit for the channel that conducts the slow delayed rectifier K⁺ current, *I_{Ks}*.

¹Department of Physiology, ²Nora Eccles Harrison Cardiovascular Research & Training Institute, ³Department of Pediatrics, University of Utah, 95 South 2000 East, Salt Lake City, Utah 84112, USA.

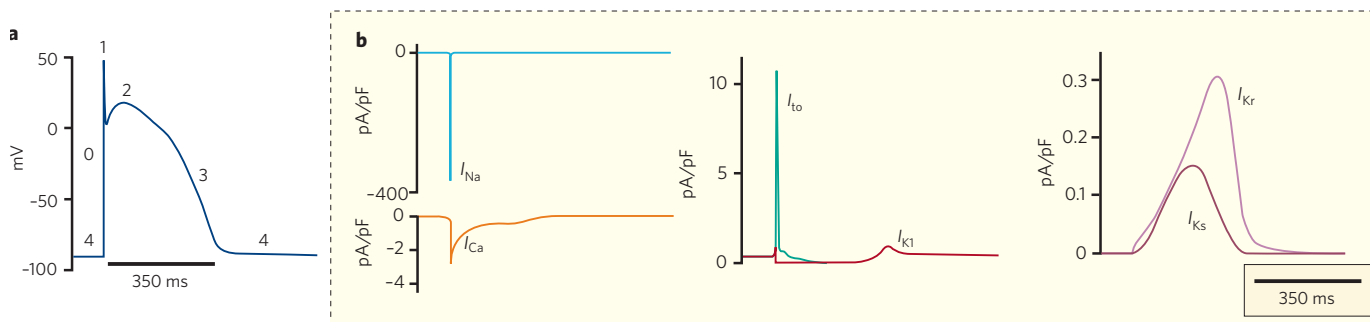


Figure 1 | Multiple ion channel currents shape the cardiac action potential. **a**, Simulated human ventricular action potential⁸⁶ with phases 0–4 indicated. An inward Na⁺ current triggers a rapid depolarization of the membrane (phase 0). Repolarization proceeds rapidly at first (phase 1), followed by a slower rate of repolarization (phase 2). The third phase ends the action potential and returns the membrane potential to the resting level (phase 4). **b**, Ion-channel currents that underlie the action potential. The rapid upstroke of the action potential (phase 0) is mediated by activation of the inward Na⁺ current, I_{Na} . Activation of the transient outward K⁺ current, I_{to} , creates a brief and partial repolarization (phase 1). The plateau of the action potential (phase 2) is prolonged owing to a maintained inward current (I_{Ca}) conducted by L-type Ca²⁺ channels and because the outward K⁺ currents are slow to activate (slow delayed rectifier K⁺ current, I_{Ks}) or exhibit strong rectification (inward rectifier K⁺ current, I_{K1} and rapid delayed rectifier K⁺ current, I_{Kr} , conducted by hERG channels). Phase 3 repolarization is mediated by the outward currents I_{Kr} and I_{K1} . The resting potential between action potentials is called phase 4. Modelling courtesy of F. Sachse, University of Utah.

(Fig. 1b). Less commonly, mutations in accessory β -subunits that coassemble with hERG or KCNQ1 α -subunits cause LQTS (Table 1). Mutations in any of the α - or β -subunits reduce outward K⁺ conductance, slow the rate of action-potential repolarization and result in electrical instability that can generate TdP.

Inherited LQTS can also result from mutations in the cardiac Na⁺ channel gene *SCN5A*. In this case, gain-of-function point mutations reduce the ability of Na⁺ channels to inactivate, resulting in a small but persistent inward current during the plateau phase that prolongs action-potential duration¹⁵. Mutations in certain LQTS genes can cause problems beyond an increased risk of arrhythmia. For example, autosomal recessive mutations in *KCNQ1*, or its accessory β -subunit *KCNE1*, cause a rare and more severe form of LQTS (Jervell and Lange-Nielsen syndrome) associated with sensorineural deafness. Loss-of-function mutations in the inward rectifier K⁺ channel (I_{K1}) gene *KCNJ2* cause Andersen–Tawil syndrome¹⁶, which is characterized by skeletal muscle periodic paralysis and ventricular arrhythmias that result from a delay in the final stage of ventricular repolarization. In most forms of LQTS, the heart is structurally normal. An exception is Timothy syndrome, a very rare and complex multi-organ disorder that is associated with congenital heart disease, extreme QT prolongation, severe and often lethal arrhythmia, syndactyly (where toes and fingers are partly fused), immune deficiency and autism¹⁷. Timothy syndrome is caused by point mutations in the cardiac L-type Ca²⁺-channel gene *CACNA1C* that disrupt channel inactivation. A summary of the ion-channel genes associated with inherited arrhythmia is presented in Table 1.

Approximately 200 LQTS-associated mutations in *HERG* have been

described (see the Gene Connection for the Heart website: <http://pc4.fsm.it:81/cardmoc/>). The functional consequence of most hERG mutations is a disruption of the folding of subunits and of trafficking of the channel to the cell surface membrane^{18,19}. Mutated and misfolded hERG subunits are usually retained in the endoplasmic reticulum in the core-glycosylated form and are rapidly degraded by the ubiquitin–proteasome pathway²⁰. Mutations can also alter hERG gating or cause dominant-negative suppression when mutant and wild-type subunits coassemble^{21,22}. An example of altered gating is the enhanced inactivation caused by point mutations located in the extracellular linker between the S5 domain and the pore helix²³. Regardless of the molecular mechanism, all LQTS-associated hERG mutations reduce current magnitude, a loss of function. By contrast, one mutation in hERG, Asn588Lys, abolishes inactivation and increases outward repolarizing current, a gain of function. This mutation hastens cardiac repolarization, shortens the QT interval and can cause ventricular fibrillation and sudden death²⁴. The finding that both loss- and gain-of-function mutations in hERG can cause lethal arrhythmia emphasizes that normal electrical activity of the heart requires a finely balanced expression of ion channels.

Structure of the hERG channel

Most of what we know about the biophysical basis of K⁺ channel gating is derived from studies of Shaker (Kv1.1) channels. The crystal structures of bacterial K⁺ channels (KcsA^{4,25}, MthK^{5,26} and KvAP⁶) and a mammalian channel (Kv1.2^{7,27}) solved by Rod MacKinnon and colleagues at Rockefeller University have provided extraordinary insights into the structural basis of channel function. Kv channels, which include

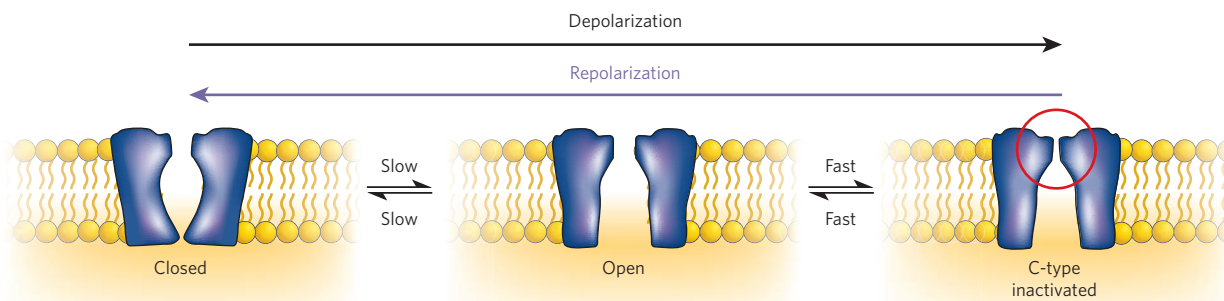


Figure 2 | Conformation of a single hERG channel is voltage dependent. Single hERG channels are either closed, open or inactivated, depending on transmembrane voltage. Channels are closed at negative voltages. Membrane depolarization slowly activates (opens) the channels, which then inactivate rapidly, especially at higher potentials. Repolarization of the membrane reverses the transitions between these channels states. C-type inactivation is thought to be caused by constriction of the selectivity filter (circled in red).

Table 1 | Ion-channel genes associated with cardiac arrhythmia

<i>HERG</i> ^{2,24} (<i>KCNH2</i>)	Romano-Ward (autosomal dominant) long QT syndrome Short QT syndrome	<i>hERG</i> (KV11.1)	<i>I_{Kr}</i>	Loss of function (decreased current) Gain of function (impaired inactivation)
<i>KCNQ1</i> ^{24,81}	Romano-Ward and Jervell and Lange-Nielsen (autosomal recessive) long QT syndromes Short QT syndrome	<i>KvLQT1</i> (KV7.1) α	<i>I_{Ks}</i>	Loss of function (decreased current) Gain of function (channels remain open)
<i>KCNE1</i> ⁸²	Romano-Ward & Jervell and Lange-Nielsen long QT syndromes	<i>minK</i> β	<i>I_{Ks}</i>	Loss of function (decreased current)
<i>KCNE2</i> ³⁸	Romano-Ward long QT syndromes	<i>MiRP1</i> β	<i>I_{Kr}</i>	Loss of function (decreased current)
<i>KCNJ2</i> ^{16,87}	Andersen-Tawil syndrome Short QT syndrome	<i>Kir2.1</i> α	<i>I_{K1}</i>	Loss of function (decreased current) Gain of function
<i>SCN5A</i> ^{83,84,88}	Romano-Ward long QT syndrome Brugada syndrome; progressive conduction system disorder	<i>NaV1.5</i> α	<i>I_{Na}</i>	Gain of function (impaired inactivation) Loss of function (decreased current)
<i>CACNA1C</i> ¹⁷	Timothy syndrome	<i>CaV1.2</i> (α_{1C})	<i>I_{Ca,L}</i>	Gain of function (impaired inactivation)
<i>RYR2</i> ⁸⁵	Catecholaminergic polymorphic ventricular tachycardia	Ryanodine receptor α		Increased release of intracellular Ca^{2+}

Alternate gene or subunit names are shown in parentheses.

Shaker, KvAP, Kv1.2 and hERG, are formed by coassembly of four identical α -subunits, each containing six α -helical transmembrane domains, S1–S6 (Fig. 4a). Each subunit comprises two functionally distinct modules, one that senses transmembrane potential (S1–S4) and one that forms the K⁺-selective pore (S5–S6).

The channel pore is asymmetrical and its dimensions change when the channel gates from a closed to an open state. The extracellular end of the pore is a narrow cylinder called the K⁺-selectivity filter that is optimally constructed for conduction of K⁺ ions. The selectivity filter of Kv channels is defined by the highly conserved sequence Thr-Val-Gly-Tyr-Gly (the K⁺ signature sequence), located at the carboxy-terminal end of the pore helix⁴. In each subunit, the side-chain hydroxyl group of Thr and the carbonyl oxygen atoms of the other four residues face towards the narrow ion-conduction pathway. Together these oxygen atoms form several octahedral binding sites that coordinate dehydrated K⁺ ions arranged in a single file and separated by a single water molecule²⁵. In hERG, the Thr and Tyr residues are substituted with Ser and Phe, but it is assumed that the structure of the selectivity filter is conserved.

Below the selectivity filter, the pore widens into a water-filled region, called the central cavity, that is lined by the S6 α -helices. In the closed state, the four S6 domains criss-cross near the cytoplasmic interface to form a narrow aperture⁴ that is too small to permit entry of ions from the cytoplasm (Fig. 4b). In response to membrane depolarization, the S6 α -helices splay outwards and increase the diameter of the aperture to allow passage of ions (Fig. 4c). In the bacterial KcsA, MthK and KvAP channels, a conserved Gly residue in S6 is proposed to serve as the hinge for the activation gate²⁶. Mutation of the putative Gly hinge in hERG alters gating but does not prevent channel opening²⁸. Although Kv1–Kv4 channels also have a Gly in the same location, a different molecular hinge may mediate channel activation. In these channels, the S6 domain is proposed to hinge at a Pro-Val-Pro motif^{27,29} that is located two helical turns below the putative Gly gating hinge. In place of the Pro-Val-Pro motif, hERG has Ile-Phe-Gly. Swapping these three amino acids in hERG creates a channel that gates between the open and inactivated state but cannot close³⁰.

The amino- and carboxy-terminal regions of hERG channels are similar to domains with well-known structure. The hERG N terminus contains a PAS (Per-Arnt-Sim) domain³¹. This structure is involved in protein–protein interactions that mediate environmental sensing in prokaryotes and transcriptional regulation in eukaryotes³². It is not known whether the PAS domain in hERG has any similar functions. However, heteromultimeric channels formed by coassembly of full-length hERG and an alternatively spliced variant that removes the PAS domain produces a channel that deactivates quickly, similar to native *I_{Kr}*³³. The C terminus contains a cyclic-nucleotide-binding domain. However, binding of cAMP to this domain has a relatively minor effect on gating, causing a shift of only a few millivolts in the voltage depend-

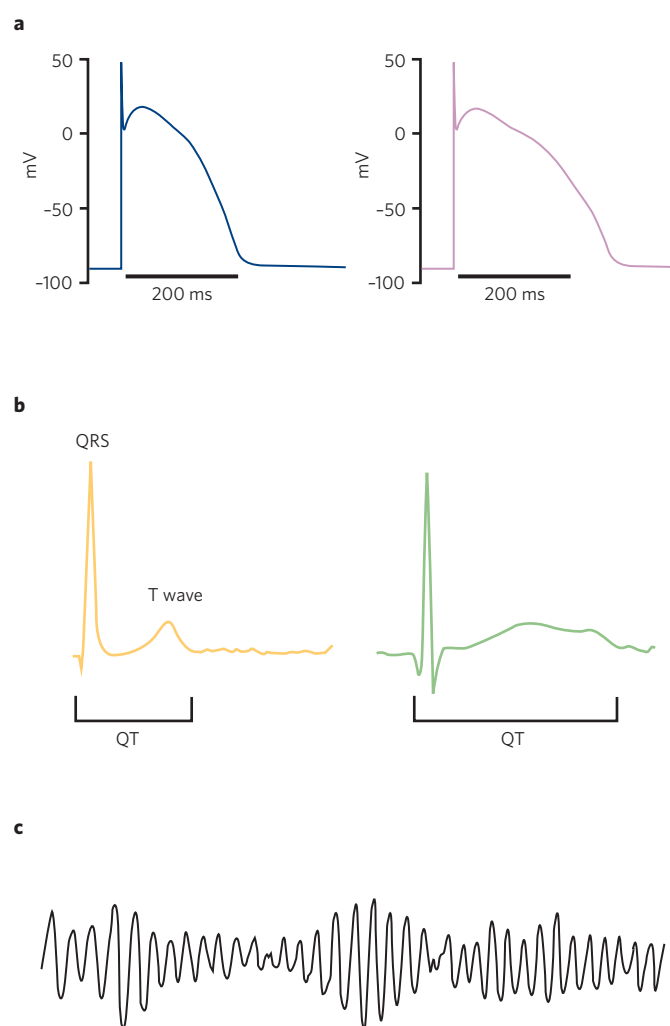


Figure 3 | Reduced hERG current delays ventricular repolarization and can induce arrhythmia. a, Normal human ventricular myocyte action potential (left) and a prolonged action potential simulated⁸⁶ by reducing hERG current by 80%. **b**, Normal ECG trace for a single cardiac cycle (left) and abnormal ECG trace with prolonged QT interval (right). The QRS complex represents depolarization, and the T wave indicates repolarization of the ventricles. The QT interval represents the time between initial depolarization and final repolarization of the ventricles. **c**, Diagram of an ECG tracing showing a *torsade de pointes* (TdP) arrhythmia.

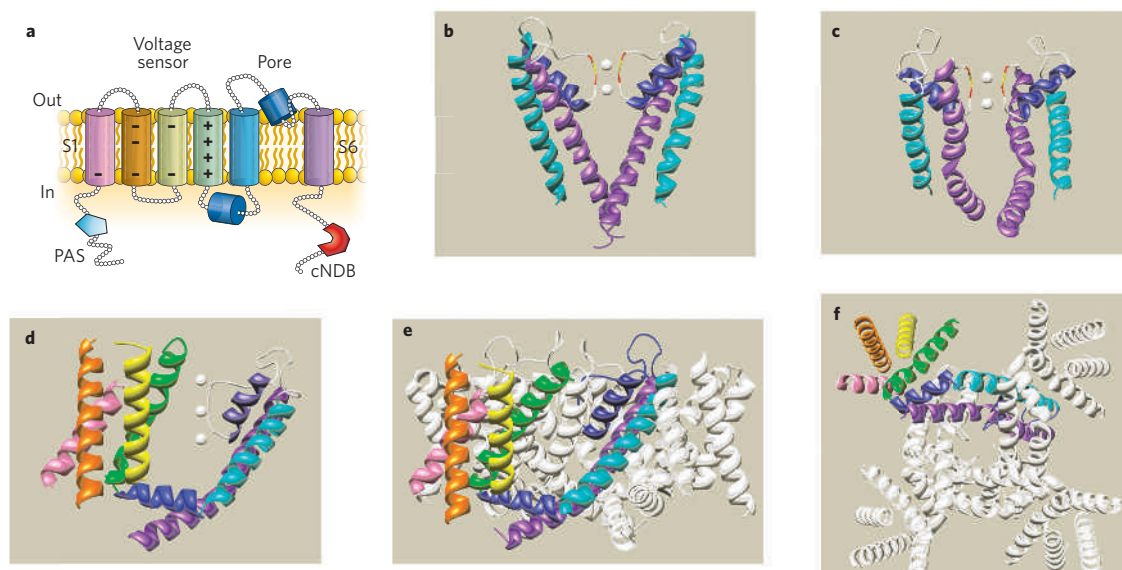


Figure 4 | Structural features of voltage-gated K^+ channels. **a**, Diagram of a single hERG subunit containing six α -helical transmembrane domains, S1–S6. Several features of the hERG subunit are highlighted, including the S4 domain, which contains multiple basic (+) amino acids, and acidic Asp residues (–) in S1–S3 that can form salt bridges with specific basic residues in S4 during gating. Also indicated are the locations of the N-terminal PAS domain and the C-terminal cyclic-nucleotide-binding domain (CNBD). **b**, Structure of a KcsA K^+ channel crystallized in the closed state⁴. Only two of the four subunits are shown. White spheres are K^+ ions located within the selectivity filter. The Gly (red) and Tyr (yellow) residues of the selectivity filter are also indicated. **c**, Structure of the pore domain of a Kv1.2 K^+ channel crystallized in the open state⁷. Only two of the four subunits are shown. **d**, Crystal structure of a single Kv1.2 α -subunit⁷ viewed from the side. Colour coding of the helical domains is the same as in panel **a**. Grey spheres represent K^+ ions. **e**, **f**, Side view (**e**) and view from the cytoplasmic side (**f**) of the membrane of the crystal structure of the complete, tetrameric Kv1.2 channel⁷.

ence of channel activation³⁴. Mutations in the cyclic-nucleotide-binding domain affect the processing of hERG channels in the endoplasmic reticulum and disrupt trafficking³⁵, but it is unclear whether this domain has an important role in gating, as it certainly does in more bona fide cyclic-nucleotide-gated (CNG) channels expressed in the rods and cones of the eye or in pacemaker (HCN) channels of the heart.

The biophysical properties of Kv channels can be altered by the interaction of pore-forming α -subunits with accessory β -subunits. For example, minK β -subunits coassemble with KCNQ1 α -subunits to form I_{Ks} channels with radically slowed activation^{36,37}. In heterologous expression systems, hERG α -subunits interact with a related β -subunit called MiRP1 (minK-related protein 1), causing decreased trafficking of channels to the cell surface and a more rapid rate of deactivation³⁸. MiRP1 is highly expressed in Purkinje fibres of the ventricular conduction system and in pacemaker cells of the atria, but its level of expression in ventricular and atrial muscle is very low^{39,40}. On the basis of this expression pattern and biophysical studies⁴¹, it is unlikely that MiRP1 interacts with hERG outside the conduction system. However, mutations³⁸ and polymorphisms⁴² in MiRP1 have been associated with ventricular arrhythmia and LQTS, suggesting a role for altered electrical activity of Purkinje fibres in these cases.

Structural basis of hERG-channel gating

The biophysical properties of I_{Kr} were first characterized in cardiac myocytes^{43–46}, but cloned hERG channels heterologously expressed in oocytes and mammalian cell lines have provided the greatest insights into the structural basis of its unusual biophysical properties. The transmembrane electrical field provides the force that drives the gating of Kv channels. The primary voltage-sensing structure of hERG and all other Kv channels is the S4 α -helical domain, which contains positively charged Lys or Arg residues in every third position. When the membrane is depolarized, the S4 moves outward. Voltage-dependent movement of the hERG S4 within the membrane electrical field can be detected as a small transient current⁴⁷ or a change in the fluorescence of a fluorophore attached to an S4 domain⁴⁸. Both techniques reveal two distinct components of voltage-sensor movement that differ by nearly

100-fold in their kinetics. The bulk of the charge movement occurs very slowly and can account for the slow rate of hERG channel activation and implies that a large energy barrier exists for transitions between the multiple closed states of the channel that are assumed to precede channel opening. Individual substitution of the six charged residues in the S4 of hERG with a bulky and hydrophobic Trp residue is well tolerated and suggests that the S4 helix is not tightly packed against neighbouring helices⁴⁹. Mutagenesis of S4 identified Arg 531 as the most important positively charged residue for proper voltage sensing in hERG^{49–51}. Arg 531 corresponds to the fourth basic residue in the S4 of Shaker, and it is proposed to move completely across the transmembrane electric field during the gating process⁵². Negatively charged acidic residues in S1–S3 form transient salt bridges with specific basic residues in the S4 to stabilize the closed, intermediate and open states of EAG (*ether-a-go-go*)⁵³ and hERG⁵⁴ channels. The outermost acidic Asp residues in S2 and S3 form a coordination site for external divalent cations that shield against salt-bridge formation and shift the voltage dependence of hERG opening to more positive potentials⁵⁵.

Precisely how the S4 domain moves in response to changes in membrane potential is hotly debated. The bulk of evidence from biophysical studies of Shaker channels suggests that the S4 domain moves only slightly (~ 5 Å)^{56,57}. By contrast, X-ray structures and tethered biotin-accessibility studies of KvAP channels suggest that the voltage sensor moves like a paddle across a much larger distance (~ 15 – 20 Å)^{58,59}. Further biophysical studies that directly compare Shaker and KvAP, and a crystal structure of any Kv channel captured in the closed state, should settle the disagreements.

Although considerable progress has been made in defining the structural basis of the voltage sensor and the activation gate in Kv channels, the mechanism and structures responsible for coupling voltage-sensor movement to channel opening (electromechanical coupling²⁷) are less well understood. The crystal structure of Kv1.2 (ref. 7) reveals that the voltage-sensing module is linked to the pore module by the S4–S5 linker, an amphipathic α -helix that runs parallel to the membrane near the cytoplasmic interface and passes over the C-terminal portion of the S6 α -helix within the same subunit (Fig. 4d–f). MacKinnon and colleagues

proposed that the S4–S5 linker functions as a lever, driven by voltage-induced changes in S4, which pushes against the S6 helices to close the channel²⁷. Studies in a variety of voltage-gated ion channels, including hERG, support an important role for the S4–S5 linker in channel gating. A direct interaction between the S4–S5 linker and the activation gate was first suggested by attempts to convert the normally voltage-independent KcsA channel into a voltage-gated channel. Shaker/KcsA chimera channels gate in a voltage-dependent manner only if the S4–S5 linker and a portion of the activation gate is derived from Shaker⁶⁰. In hERG, an electrostatic interaction between specific residues in the S4–S5 linker (Asp 540) and the C-terminal region of the S6 domain (Arg 665) stabilizes the closed channel conformation⁶¹. In addition to helping resolve the debate over S4 movement, crystal structures of the closed and open states of a single type of Kv channel would greatly aid our understanding of the mechanism of electromechanical coupling.

Two different mechanisms of Kv-channel inactivation have been identified. Rapid ‘N-type’ inactivation takes place when a cytoplasmic ball-like structure tethered to the N terminus of a single subunit occludes the inner mouth of the channel pore. Truncation of the N-terminal domain eliminates rapid N-type inactivation in Shaker⁶² and unveils a much slower ‘C-type’ inactivation. Mutation of a single residue in the pore helix of Shaker (Thr449Val) eliminates C-type inactivation. Although inactivation of hERG channels is very rapid, it is not greatly affected by N-terminal deletion⁶³. However, inactivation of hERG is eliminated by a mutation (Ser631Val)⁶⁴ that is homologous to the Thr449Val that removes C-type inactivation in Shaker. Thus, hERG inactivates by using a C-type mechanism that is much faster than that observed for Shaker or other Kv channels. The mechanism of C-type inactivation, deduced by study of Kv channels other than hERG, is thought to be a slight constriction of the selectivity filter⁶⁵ that may only occur when the outermost K⁺-binding site is not occupied by an ion⁶⁶.

Drug-induced block of hERG channels

Common medications can prolong the QT interval and induce TdP, effects similar to those of inherited LQTS. This toxicity is widespread, as indicated by analysis of a drug-monitoring database maintained by the World Health Organization⁶⁷. Drug-induced TdP by the antiarrhythmic drug quinidine is a relatively frequent side effect, affecting 2–9% of treated patients³. However, induction of TdP by drugs other than

antiarrhythmic agents is rare. For example, cisapride-induced TdP occurred in only about 1 out of 120,000 patients prescribed this medication⁶⁸. Nonetheless, this level of risk is obviously unacceptable for drugs such as cisapride or terfenadine that are used to treat non-life-threatening gastrointestinal disorders or allergies. Recognition of rare drug-induced TdP prompted regulatory agencies to remove from the market, or relegate to restricted use, several drugs, including cisapride, sertindole, grepafloxacin, terfenadine and astemizole.

Inherited LQTS and TdP are caused by loss-of-function mutations in several cardiac K⁺ channels (Table 1). However, in clinical practice, drug-induced QT prolongation and TdP are caused by direct blockage of hERG channels⁶⁹, interference with hERG-channel trafficking to the cell surface⁷⁰ or drug–drug interactions (for example, interference with metabolism) that ultimately lead to a reduction in hERG-channel current. This raises the obvious question as to why blockage of hERG channels in particular causes drug-induced arrhythmia. As discussed below, the answer may be that hERG has structural features that can more effectively accommodate the binding of drugs compared with other K⁺ channels.

hERG channels are blocked by chemicals with diverse structures that encompass several therapeutic drug classes, including antiarrhythmic, psychiatric, antimicrobial and antihistamine. This unpredictability has frustrated conventional drug-design approaches to circumvent the arrhythmia side effect⁶⁹. However, all is not hopeless, because pharmacophore models based on a restricted and well-characterized chemical class of compound have some predictive value^{71–73}. It is now common practice for pharmaceutical companies to screen compounds for hERG-channel activity early during preclinical safety assessment⁷⁴. This approach has limitations because drug-induced blockage of hERG does not always prolong the QT interval or induce TdP. Additional activities of a drug (for example, blocking L-type Ca²⁺ channels) can counteract the effects of hERG-channel blockage on the duration of action potentials. Nonetheless, measuring hERG current is more suitable for moderate to high-throughput screening than assessing changes in QT interval and arrhythmia risk in animals. The recent development of high-capacity voltage-clamp instruments based on planar patch technology⁷⁵ promises to streamline these screening efforts.

The hERG channel is unusually susceptible to blockage by drugs in comparison with other Kv channels, suggesting that it has a unique binding site. An Ala-scanning mutagenesis approach was used to identify residues of hERG that interact with several drugs. Residues within the pore module were individually mutated to an Ala, and the resulting mutant channels assayed for sensitivity to potent hERG blockers²⁸. Mutation of two polar residues (Thr 623 and Ser 624) located at the base of the pore helix and two aromatic residues (Tyr 652 and Phe 656) located in the S6 domain of the hERG subunit decreased the affinity of MK-499, a potent antiarrhythmic drug. The same residues were found to be important for binding of cisapride, terfenadine and several other drugs from diverse chemical and therapeutic classes⁷⁴. The side chains of all four residues are orientated towards the large central cavity of the channel (Fig. 5), which is consistent with the observation that hERG channels are only blocked by these drugs after the channel has opened. The two pore helix residues (Thr 623 and Ser 624) are highly conserved in Kv channels and thus cannot easily explain the promiscuous blocking by drugs of hERG. However, the two S6 residues (Tyr 652 and Phe 656) are not conserved — most Kv channels have an Ile and a Val in homologous positions. Further mutagenesis identified the most relevant physicochemical properties of these two S6 residues. Potent blockage of hERG by cisapride and terfenadine requires an aromatic residue in position 652, suggesting the possible importance of a cation– π interaction between the positively charged N of the drug and the π -electrons of Tyr 652. A strongly hydrophobic attraction with Phe656 was identified for the same drugs³⁰. *In silico* docking of drugs to homology models of hERG supports these conclusions^{76–78}. Perhaps the multiple aromatic side chains (eight per channel), arranged in two concentric rings, can accommodate multiple and compound-specific interactions, partially explaining the surprising chemical diversity of hERG blockers.



Figure 5 | Model of the hERG drug-binding site. Homology model of the hERG-channel pore module (S5, S6 and pore helix of two subunits) based on the crystal structure of KvAP (ref. 6). The key residues that interact with structurally diverse drugs are highlighted, including Thr 623 (orange) and Ser 624 (white) located at the base of the pore helix, and Tyr 652 (yellow) and Phe 656 (magenta) located on the S6 domain.

Future directions

hERG has received a tremendous amount of attention since its discovery in 1994 because inherited mutations or drug-induced blockade of channels increases the risk of lethal arrhythmia. Despite intense scrutiny, many issues regarding the physiological functions and mechanisms of gating of hERG channels remain unresolved. Compared with their well-understood role in cardiac repolarization, the physiological functions of hERG channels in the central nervous system are poorly understood. Mutations in hERG may cause subtle neural abnormalities, previously overlooked because the cardiovascular phenotype was so severe. hERG-channel expression is upregulated in tumour cells^{10,79}, but it remains to be determined whether this is a key event or just one of many changes in gene expression that accompanies cell proliferation.

Current treatment for inherited LQTS consists of β -adrenergic blockers, followed by implantable defibrillators if drug therapy is inadequate. Screening efforts to avoid drug blockage of hERG led to the recent discovery of novel compounds that, ironically, increase channel activity⁸⁰. Studies are needed to determine whether activators can safely counteract the consequences of reduced hERG-channel current in LQTS.

Although the crucial residues of the drug-binding site seem to have been identified, clearly not all drugs interact with hERG in the same manner. Further studies are needed to explain why hERG in particular is blocked by so many different drugs. Regardless of the mechanisms responsible for this promiscuity, a combination of ligand-docking and receptor homology models and quantitative structure–activity relationship analyses provides the best hope of truly predictive *in silico* assessment of the potential hERG-binding affinity of new chemical entities. ■

- Warmke, J. W. & Ganetzky, B. A family of potassium channel genes related to *eag* in *Drosophila* and mammals. *Proc. Natl Acad. Sci. USA* **91**, 3438–3442 (1994).
- Curran, M. E. *et al.* A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome. *Cell* **80**, 795–803 (1995).
- Haverkamp, W. *et al.* The potential for QT prolongation and proarrhythmia by non-antiarrhythmic drugs: clinical and regulatory implications. Report on a policy conference of the European Society of Cardiology. *Eur. Heart J.* **21**, 1216–1231 (2000).
- Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* **280**, 69–77 (1998).
- Jiang, Y. *et al.* Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**, 515–522 (2002).
- Jiang, Y. *et al.* X-ray structure of a voltage-dependent K^+ channel. *Nature* **423**, 33–41 (2003).
- Long, S. B., Campbell, E. B. & MacKinnon, R. Crystal structure of a mammalian voltage-dependent Shaker family K^+ channel. *Science* **309**, 897–903 (2005).
- Farrelly, A. M. *et al.* Expression and function of KCNH2 (HERG) in the human jejunum. *Am. J. Physiol.* **284**, G883–G895 (2003).
- Chiesa, N., Rosati, B., Arcangeli, A., Olivetto, M. & Wanke, E. A novel role for HERG K^+ channels: spike-frequency adaptation. *J. Physiol. (Lond.)* **501**, 313–318 (1997).
- Smith, G. A. *et al.* Functional up-regulation of HERG K^+ channels in neoplastic hematopoietic cells. *J. Biol. Chem.* **277**, 18528–18534 (2002).
- Shi, W. *et al.* Identification of two nervous system-specific members of the *erg* potassium channel gene family. *J. Neurosci.* **17**, 9423–9432 (1997).
- Wimmers, S., Wulfsen, I., Bauer, C. K. & Schwarz, J. R. Erg1, erg2 and erg3 K channel subunits are able to form heteromultimers. *Pflügers Arch.* **441**, 450–455 (2001).
- Sanguinetti, M. C., Jiang, C., Curran, M. E. & Keating, M. T. A mechanistic link between an inherited and an acquired cardiac arrhythmia: HERG encodes the I_{Kr} potassium channel. *Cell* **81**, 299–307 (1995).
- Trudeau, M., Warmke, J. W., Ganetzky, B. & Robertson, G. A. HERG, a human inward rectifier in the voltage-gated potassium channel family. *Science* **269**, 92–95 (1995).
- Bennett, P. B., Yazawa, K., Makita, N. & George, A. L. Molecular mechanism for an inherited cardiac arrhythmia. *Nature* **376**, 683–685 (1995).
- Plaster, N. M. *et al.* Mutations in Kir2.1 cause the developmental and episodic electrical phenotypes of Andersen's syndrome. *Cell* **105**, 511–519 (2001).
- Splawski, I. *et al.* $Ca_v1.2$ calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* **119**, 19–31 (2004).
- Furutani, M. *et al.* Novel mechanism associated with an inherited cardiac arrhythmia: defective protein trafficking by the mutant HERG (G601S) potassium channel. *Circulation* **99**, 2290–2294 (1999).
- Delisle, B. P., Anson, B. D., Rajamani, S. & January, C. T. Biology of cardiac arrhythmias: ion channel protein trafficking. *Circ. Res.* **94**, 1418–1428 (2004).
- Gong, Q., Keeney, D. R., Molinari, M. & Zhou, Z. Degradation of trafficking-defective long QT syndrome type II mutant channels by the ubiquitin–proteasome pathway. *J. Biol. Chem.* **280**, 19419–19425 (2005).
- Sanguinetti, M. C., Curran, M. E., Spector, P. S. & Keating, M. T. Spectrum of HERG K^+ channel dysfunction in an inherited cardiac arrhythmia. *Proc. Natl Acad. Sci. USA* **93**, 2208–2212 (1996).
- Kagan, A., Yu, Z., Fishman, G. I. & McDonald, T. V. The dominant negative LQT2 mutation A561V reduces wild-type HERG expression. *J. Biol. Chem.* **275**, 11241–11248 (2000).
- Nakajima, T. *et al.* Novel mechanism of HERG current suppression in LQT2: shift in voltage dependence of HERG inactivation. *Circ. Res.* **83**, 415–422 (1998).
- Brugada, R. *et al.* Sudden death associated with short-QT syndrome linked to mutations in HERG. *Circulation* **109**, 30–35 (2004).
- Zhou, Y., Morais-Cabral, J. H., Kaufman, A. & MacKinnon, R. Chemistry of ion coordination and hydration revealed by a K^+ channel–Fab complex at 2.0 Å resolution. *Nature* **414**, 43–48 (2001).
- Jiang, Y., Lee, A., Cadene, M., Chait, B. T. & MacKinnon, R. The open pore conformation of potassium channels. *Nature* **417**, 523–526 (2002).
- Long, S. B., Campbell, E. B. & MacKinnon, R. Voltage sensor of Kv1.2: structural basis of electromechanical coupling. *Science* **309**, 903–908 (2005).
- Mitcheson, J. S., Chen, J., Lin, M., Culberson, C. & Sanguinetti, M. C. A structural basis for drug-induced long QT syndrome. *Proc. Natl Acad. Sci. USA* **97**, 12329–12333 (2000).
- Webster, S. M., Del Camino, D., Dekker, J. P. & Yellen, G. Intracellular gate opening in Shaker K^+ channels defined by high-affinity metal bridges. *Nature* **428**, 864–868 (2004).
- Fernandez, D., Ghanta, A., Kauffman, G. W. & Sanguinetti, M. C. Physicochemical features of the hERG channel drug binding site. *J. Biol. Chem.* **279**, 10120–10127 (2004).
- Morais Cabral, J. H. *et al.* Crystal structure and functional analysis of the HERG potassium channel N terminus: a eukaryotic PAS domain. *Cell* **95**, 649–655 (1998).
- Taylor, B. L. & Zhulin, I. B. PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol. Mol. Biol. Rev.* **63**, 479–506 (1999).
- London, B. *et al.* Two isoforms of the mouse *ether-a-go-go*-related gene coassemble to form channels with properties similar to the rapidly activating component of the cardiac delayed rectifier K^+ current. *Circ. Res.* **81**, 870–878 (1997).
- Cui, J., Melman, Y., Palma, E., Fishman, G. I. & McDonald, T. V. Cyclic AMP regulates the HERG K^+ channel by dual pathways. *Curr. Biol.* **10**, 671–674 (2000).
- Akhavan, A. *et al.* Identification of the cyclic-nucleotide-binding domain as a conserved determinant of ion-channel cell-surface localization. *J. Cell Sci.* **118**, 2803–2812 (2005).
- Barhanin, J. *et al.* KvLTQ1 and Isk (minK) proteins associate to form the I_{Ks} cardiac potassium channel. *Nature* **384**, 78–80 (1996).
- Sanguinetti, M. C. *et al.* Coassembly of KvLTQ1 and minK (IsK) proteins to form cardiac I_{Ks} potassium channel. *Nature* **384**, 80–83 (1996).
- Abbott, G. W. *et al.* MiRP1 forms I_{Kr} potassium channels with HERG and is associated with cardiac arrhythmia. *Cell* **97**, 175–187 (1999).
- Pourrier, M., Zicha, S., Ehrlich, J., Han, W. & Nattel, S. Canine ventricular KCNE2 expression resides predominantly in Purkinje fibers. *Circ. Res.* **93**, 189–191 (2003).
- Lundquist, A. L. *et al.* Expression of multiple KCNE genes in human heart may enable variable modulation of I_{Ks} . *J. Mol. Cell. Cardiol.* **38**, 277–287 (2005).
- Weerapura, M., Nattel, S., Chartier, D., Caballero, R. & Hebert, T. E. A comparison of currents carried by HERG, with and without coexpression of MiRP1, and the native rapid delayed rectifier current. Is MiRP1 the missing link? *J. Physiol. (Lond.)* **540**, 15–27 (2002).
- Sesti, F. *et al.* A common polymorphism associated with antibiotic-induced cardiac arrhythmia. *Proc. Natl Acad. Sci. USA* **97**, 10613–10618 (2000).
- Noble, D. & Tsien, R. W. Outward membrane currents activated in the plateau range of potentials in cardiac Purkinje fibres. *J. Physiol. (Lond.)* **200**, 205–231 (1969).
- Shibasaki, T. Conductance and kinetics of delayed rectifier potassium channels in nodal cells of the rabbit heart. *J. Physiol. (Lond.)* **387**, 227–250 (1987).
- Clay, J. R., Hill, C. E., Roitman, D. & Shrier, A. Repolarization current in embryonic chick atrial heart cells. *J. Physiol. (Lond.)* **403**, 525–537 (1988).
- Sanguinetti, M. C. & Jurkiewicz, N. K. Two components of cardiac delayed rectifier K^+ current: Differential sensitivity to block by class III antiarrhythmic agents. *J. Gen. Physiol.* **96**, 195–215 (1990).
- Piper, D. R., Varghese, A., Sanguinetti, M. C. & Tristani-Firouzi, M. Gating currents associated with intramembrane charge displacement in HERG potassium channels. *Proc. Natl Acad. Sci. USA* **100**, 10534–10539 (2003).
- Smith, P. L. & Yellen, G. Fast and slow voltage sensor movements in HERG potassium channels. *J. Gen. Physiol.* **119**, 275–293 (2002).
- Subbiah, R. N., Kondo, M., Campbell, T. J. & Vandenberg, J. I. Tryptophan scanning mutagenesis of the HERG K^+ channel: The S4 domain is loosely packed and likely to be lipid exposed. *J. Physiol. (Lond.)* **569**, 367–379 (2005).
- Piper, D. R., Hinz, W. A., Talluri, C. K., Sanguinetti, M. C. & Tristani-Firouzi, M. Regional specificity of human *ether-a-go-go*-related gene channel activation and inactivation gating. *J. Biol. Chem.* **280**, 7206–7217 (2005).
- Zhang, M., Liu, J. & Tseng, G. N. Gating charges in the activation and inactivation processes of the HERG channel. *J. Gen. Physiol.* **124**, 703–718 (2004).
- Larsson, H. P., Baker, O. S., Dhillon, D. S. & Isacoff, E. Y. Transmembrane movement of the Shaker K^+ channel S4. *Neuron* **16**, 387–397 (1996).
- Silverman, W. R., Roux, B. & Papazian, D. M. Structural basis of two-stage voltage-dependent activation in K^+ channels. *Proc. Natl Acad. Sci. USA* **100**, 2935–2940 (2003).
- Liu, J., Zhang, M., Jiang, M. & Tseng, G. N. Negative charges in the transmembrane domains of the HERG K channel are involved in the activation- and deactivation-gating processes. *J. Gen. Physiol.* **121**, 599–614 (2003).
- Fernandez, D., Ghanta, A., Kinard, K. I. & Sanguinetti, M. C. Molecular mapping of a site for Cd²⁺-induced modification of human *ether-a-go-go*-related gene (hERG) channel activation. *J. Physiol. (Lond.)* **567**, 737–755 (2005).
- Laine, M., Papazian, D. M. & Roux, B. Critical assessment of a proposed model of Shaker. *FEBS Lett.* **564**, 257–263 (2004).
- Posson, D. J., Ge, P., Miller, C., Bezanilla, F. & Selvin, P. R. Small vertical movement of a K^+ channel voltage sensor measured with luminescence energy transfer. *Nature* **436**, 848–851 (2005).
- Jiang, Y., Ruta, V., Chen, J., Lee, A. & MacKinnon, R. The principle of gating charge movement in a voltage-dependent K^+ channel. *Nature* **423**, 42–48 (2003).
- Ruta, V., Chen, J. & MacKinnon, R. Calibrated measurement of gating-charge arginine displacement in the KvAP voltage-dependent K^+ channel. *Cell* **123**, 463–475 (2005).
- Lu, Z., Klem, A. M. & Ramu, Y. Coupling between voltage sensors and activation gate in voltage-gated K^+ channels. *J. Gen. Physiol.* **120**, 663–676 (2002).
- Tristani-Firouzi, M., Chen, J. & Sanguinetti, M. C. Interactions between S4–S5 linker and S6 transmembrane domain modulate gating of HERG K^+ channels. *J. Biol. Chem.* **277**, 18994–19000 (2002).
- Spector, P. S., Curran, M. E., Zou, A., Keating, M. T. & Sanguinetti, M. C. Fast inactivation

- causes rectification of the I_{Kr} channel. *J. Gen. Physiol.* **107**, 611–619 (1996).
63. Smith, P. L., Baukrowitz, T. & Yellen, G. The inward rectification mechanism of the HERG cardiac potassium channel. *Nature* **379**, 833–836 (1996).
 64. Fan, J.-S., Jinag, M., Dun, W., McDonald, T. V. & Tseng, G.-N. Effects of outer mouth mutations on hERG channel function: A comparison with similar mutations in the Shaker channel. *Biophys. J.* **76**, 3128–3140 (1999).
 65. Baukrowitz, T. & Yellen, G. Modulation of K^+ current by frequency and external $[K^+]$: A tale of two inactivation mechanisms. *Neuron* **15**, 951–960 (1995).
 66. Kiss, L. & Korn, S. J. Modulation of C-type inactivation by K^+ at the potassium channel selectivity filter. *Biophys. J.* **74**, 1840–1849 (1998).
 67. De Bruin, M. L., Pettersson, M., Meyboom, R. H., Hoes, A. W. & Leufkens, H. G. Anti-HERG activity and the risk of drug-induced arrhythmias and sudden death. *Eur. Heart J.* **26**, 590–597 (2005).
 68. Vitola, J., Vukanovic, J. & Roden, D. M. Cisapride-induced torsades de pointes. *J. Cardiovasc. Electrophysiol.* **9**, 1109–1113 (1998).
 69. Recanatini, M., Poluzzi, E., Masetti, M., Cavalli, A. & De Ponti, F. QT prolongation through hERG K^+ channel blockade: current knowledge and strategies for the early prediction during drug development. *Med. Res. Rev.* **25**, 133–166 (2005).
 70. Kuryshv, Y. A. *et al.* Pentamidine-induced long QT syndrome and block of hERG trafficking. *J. Pharmacol. Exp. Ther.* **312**, 316–323 (2005).
 71. Cavalli, A., Poluzzi, E., De Ponti, F. & Recanatini, M. Toward a pharmacophore for drugs inducing the long QT syndrome: Insights from a CoMFA study of HERG K^+ channel blockers. *J. Med. Chem.* **45**, 3844–3853 (2002).
 72. Ekins, S., Crumb, W. J., Sarazan, R. D., Wikel, J. H. & Wrighton, S. A. Three-dimensional quantitative structure–activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *J. Pharmacol. Exp. Ther.* **301**, 427–434 (2002).
 73. Pearlstein, R. A. *et al.* Characterization of HERG potassium channel inhibition using CoMSIA 3D QSAR and homology modeling approaches. *Bioorg. Med. Chem. Lett.* **13**, 1829–1835 (2003).
 74. Sanguinetti, M. C. & Mitcheson, J. S. Predicting drug-hERG channel interactions that cause acquired long QT syndrome. *Trends Pharmacol. Sci.* **26**, 119–124 (2005).
 75. Dubin, A. E. *et al.* Identifying modulators of hERG channel activity using the PatchXpress planar patch clamp. *J. Biomol. Screen.* **10**, 168–181 (2005).
 76. Pearlstein, R., Vaz, R. & Rampe, D. Understanding the structure–activity relationship of the human ether-a-go-go-related gene cardiac K^+ channel. A model for bad behavior. *J. Med. Chem.* **46**, 2017–2022 (2003).
 77. Rajamani, R., Tounge, B. A., Li, J. & Reynolds, C. H. A two-state homology model of the hERG K^+ channel: application to ligand binding. *Bioorg. Med. Chem. Lett.* **15**, 1737–1741 (2005).
 78. Osterberg, F. & Aqvist, J. Exploring blocker binding to a homology model of the open hERG K^+ channel using docking and molecular dynamics methods. *FEBS Lett.* **579**, 2939–2944 (2005).
 79. Lastraioli, E. *et al.* *herg1* gene and HERG1 protein are overexpressed in colorectal cancers and regulate cell invasion of tumor cells. *Cancer Res.* **64**, 606–611 (2004).
 80. Kang, J. *et al.* Discovery of a small molecule activator of the Human Ether-a-go-go-Related Gene (HERG) cardiac K^+ channel. *Mol. Pharmacol.* **67**, 827–836 (2005).
 81. Wang, Q. *et al.* Positional cloning of a novel potassium channel gene: KVLQT1 mutations cause cardiac arrhythmias. *Nature Genetics* **12**, 17–23 (1996).
 82. Splawski, I., Tristani-Firouzi, M., Lehmann, M. H., Sanguinetti, M. C. & Keating, M. T. Mutations in the *hminK* gene cause long QT syndrome and suppress I_{Ks} function. *Nature Genet.* **17**, 338–340 (1997).
 83. Wang, Q. *et al.* SCN5A mutations associated with an inherited cardiac arrhythmia, long QT syndrome. *Cell* **80**, 805–811 (1995).
 84. Chen, Q. *et al.* Genetic basis and molecular mechanism for idiopathic ventricular fibrillation. *Nature* **392**, 293–296 (1998).
 85. Priori, S. G. *et al.* Mutations in the cardiac ryanodine receptor gene (hRyR2) underlie catecholaminergic polymorphic ventricular tachycardia. *Circulation* **103**, 196–200 (2001).
 86. Priebe, L. & Beuckelmann, D. J. Simulation study of cellular electric properties in heart failure. *Circ. Res.* **82**, 1206–1223 (1998).
 87. Priori, S. G. *et al.* A novel form of short QT syndrome (SQT3) is caused by a mutation in the *KCNJ2* gene. *Circ. Res.* **96**, 800–807 (2005).
 88. Schott, J.-J. *et al.* Cardiac conduction defects associate with mutations in *SCN5A*. *Nature Genet.* **23**, 20–21 (1999).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to M.C.S. (sanguinetti@cvrti.utah.edu).

K_{ATP} channels as molecular sensors of cellular metabolism

Colin G. Nichols¹

In responding to cytoplasmic nucleotide levels, ATP-sensitive potassium (K_{ATP}) channel activity provides a unique link between cellular energetics and electrical excitability. Over the past ten years, a steady drumbeat of crystallographic and electrophysiological studies has led to detailed structural and kinetic models that define the molecular basis of channel activity. In parallel, the uncovering of disease-causing mutations of K_{ATP} has led to an explanation of the molecular basis of disease and, in turn, to a better understanding of the structural basis of channel function.

Action potentials, which define electrical excitability, depend on the activity of voltage-gated Na⁺, Ca²⁺ and K⁺ (Kv) channels (see the review in this issue by Sanguinetti and Tristani-Firouzi, p. 463). But electrical activity without modulation would be like an automobile with one speed, no steering and no brakes. In addition to the channels that play the leading parts, bio-electric signalling involves many crucial bit players (see the review in this issue by Miller, p. 484) that provide essential gears, steering and brakes. These include a family of K⁺ channels, the so-called inward rectifier (Kir) channels, which lack steep voltage-gating. By stabilizing the resting membrane potential, these channels act as a brake on excitability. One member, the K_{ATP} channel, modulates electrical activity in multiple tissues^{1–4}. In a complex interplay of mechanisms, the K_{ATP} channel is inhibited by the non-hydrolytic binding of ATP, but activated by interactions with Mg²⁺-bound nucleotides (Mg-nucleotides) at separate sites. The inhibitory effect dominates, and channels are closed, when cellular phosphorylation potential is high, but as metabolism decreases, the activating effect wins out and channels open. In this way, the channel provides a unique electrical transducer of the metabolic state of the cell.

What mechanisms of channel regulation allow this transduction? Many questions remain to be addressed, but as discussed below, the cloning of constituent subunits of the K_{ATP} channel and the crystallization of structurally related proteins have provided the raw materials to gain a detailed understanding of how nucleotides interact with the K_{ATP} channel and an explanation of the molecular basis of channel activity and of channel-dependent human disease.

Protein architecture of K_{ATP} channels

K_{ATP} channels are formed by the unique combination of two dissimilar proteins² (Fig. 1a). One, Kir6, is a member of the Kir channel family; the other, SUR (for sulphonylurea receptor), is a member of the ATP-binding cassette (ABC) protein family, which includes CFTR (cystic fibrosis transmembrane conductance regulator; see the review in this issue by Gadsby, p. 477). As discussed in detail below, ATP inhibition results from interaction with the Kir6 subunits, whereas Mg-nucleotide activation reflects interaction with SUR subunits. In the vertebrate genome, there are two Kir6 genes (*KCNJ8*, *KIR6.1* and *KCNJ11*, *KIR6.2*) and two SUR genes (*ABCC8*, *SUR1* and *ABCC9*, *SUR2*), and co-expression of combinations of the two genes replicates all of the key features of classical K_{ATP} channel activity^{5–7}. Biophysical and biochemical experiments established that four Kir6.2 subunits generate the channel pore, each Kir6.2

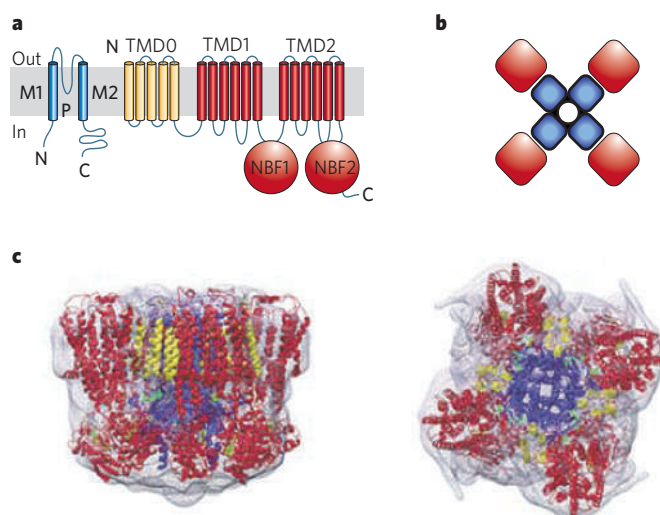


Figure 1 | The K_{ATP} channel is formed from two dissimilar subunits.

a, Inward rectifier K⁺ channel Kir6 subunits generate the channel pore and sulphonylurea receptor (SUR) subunits generate the regulatory subunit. TMD, transmembrane domain; NBF, nucleotide-binding fold; M1, M2, transmembrane helices; P, pore. **b**, The channel is a functional octamer of four Kir6 subunits, and each subunit is associated with four SUR subunits. **c**, Images at 18 Å resolution of the entire K_{ATP} complex viewed in the plane of the membrane (left) or from above the membrane (right) require tight packing of subunit models. Reproduced, with permission, from ref. 33.

being associated with one SUR1 protein^{8,9} (Fig. 1b). Stoichiometry has not been tested for other Kir6/SUR combinations, and it is conceivable that Kir6 complexes occur naturally without an associated SUR subunit, or with different stoichiometry. In recombinant cells, Kir6.2 can form channels without SUR1 when truncated at the carboxyl terminus¹⁰. In this case, truncation removes a trafficking sequence (Arg-Lys-Arg) that otherwise causes Kir6.2 retention in the endoplasmic reticulum (ER). Similarly, SUR1, which also contains an ER-retention sequence, is not efficiently trafficked to the surface membrane in the absence of Kir6.2, which indicates that the subunits normally act to mutually shield these trafficking signals¹⁰.

¹Department of Cell Biology and Physiology, Washington University School of Medicine, 660 South Euclid Avenue, St Louis, Missouri 63110, USA.

A high-resolution model of Kir6.2

Crystallization of the bacterial KcsA K⁺ channel¹¹ heralded a revolution in our understanding of channel structure. The overall architecture of the KcsA transmembrane domain, consisting of two transmembrane helices (M1 and M2) bridged by an extracellular loop that generates the narrow portion of the pore and controls ion selectivity, is likely to be present in all major cation channel family members, including Kir6.2 (Fig. 2a). Crystallization of linked amino- and carboxy-terminal domains of Kir3.1 (ref. 12) and Kir2.1 (ref. 13), and of full-length prokaryotic Kir channel homologues KirBac1.1 (ref. 14) and KirBac3.1 (ref. 15) (Protein Data Bank accession numbers 1XL4 and 1XL6), revealed an extended cytoplasmic pore and the likely location of ligand-binding sites. A consensus model of Kir6.2, based on elements of both the eukaryotic and the prokaryotic Kir structures^{16,17}, provides satisfying agreement with mutational studies of Kir6.2, particularly with respect to the inhibitory ATP-binding site (Fig. 2b; see below). At the membrane interface, the verity of the model, which relies heavily on the KirBac1.1 structure, is less clear. KirBac1.1 and KirBac3.1 are part of a distinct subfamily of prokaryotic Kir proteins, with potentially important primary sequence differences and functional differences from eukaryotic Kir channels^{18,19} that need to be understood for the correct interpretation of molecular models. The lesson from CLC chloride channels (see p. 484), that homologous structures can encode both ion channels and ion exchangers, is a sobering one. Importantly, although all eukaryotic Kir channels are strongly activated by interaction with phosphatidylinositol-4,5-bisphosphate (PIP₂), KirBac1.1 is inhibited by PIP₂ (ref. 19). The linkers between the cytoplasmic domain and the transmembrane domain are shorter in KirBac1.1 than in euk-Kir channels¹⁹ and the missing residues in the C terminus include two key residues (Arg 176 and Arg 177 in Kir6.2) that are predicted to interact directly with PIP₂ (refs 20–22). It is an intriguing hypothesis that KirBac channels are normally active in bacterial membranes (which typically contain no phosphoinositides) and that eukaryotic Kir channels acquired these additional residues specifically to restore function in the face of otherwise inhibitory PIP₂.

Templates for SUR

As with other members of ABC subfamily C (see p. 477), SUR1 (ABCC8) and SUR2 (ABCC9) each contain two six-helix transmembrane domains (TMD1 and TMD2). In addition, SURs have an N-terminal TMD0 domain that consists of five transmembrane helices (Fig. 1a). There is evidence²³ that TMD0 is crucial for trafficking Kir6.2 subunits to the surface membrane, and its role in controlling the gating of Kir6.2 (refs 24, 25) suggests an intimate relationship with the pore-forming subunit. Low homology between the transmembrane domains of crystallized ABC proteins and SURs makes molecular modelling risky, although provocative insights and hypotheses have been gleaned from the exercise^{2,26}, including speculations about locations of the binding sites for the channel opener diazoxide and the inhibitory sulphonylurea glibenclamide².

SUR also contains one nucleotide-binding fold (NBF1) between TMD1 and TMD2, and a second (NBF2) after TMD2 (Fig. 1a). Several crystal structures are now available for isolated NBFs from bacterial ABC proteins^{27–29}; in almost every case, the NBFs crystallize as 'head-to-tail' dimers. Rather than each NBF having a self-contained nucleotide-binding site (ABS), each ABS is formed at the dimer interface and contains elements from both NBFs. CFTR mutant cycle analysis provides evidence for NBF1–NBF2 interactions in the full-length protein (see p. 477), and modelling of the SUR NBFs as heterodimers²⁶ (Fig. 3b) seems reasonable given the high degree of conservation among NBF sequences. Experimentally, azido-ATP labelling experiments have shown that nucleotide binding to NBF2 stabilizes ATP binding to NBF1 (ref. 30). In addition, NBF1 and NBF2 can be co-immunoprecipitated³¹, and a soluble NBF1–GFP (green fluorescent protein) construct is recruited to the membrane by a TMD2–NBF2 construct when co-expressed in insect cells³².

Recently, complete K_{ATP} complexes generated from tandem SUR1–

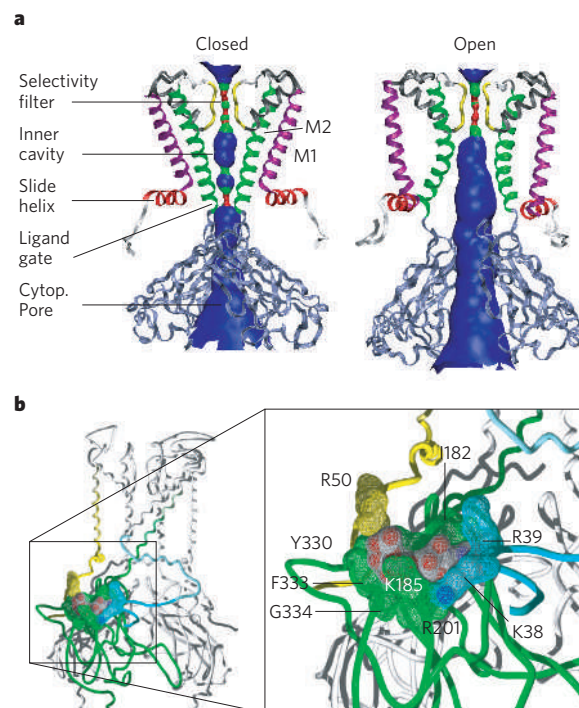


Figure 2 | ATP-dependent gating in high definition. **a**, A model inward rectifier K⁺ (Kir) channel showing potential gating states. Closed and open structures modelled on two-dimensional crystals of KirBac3.1 are shown in two different conformational states¹⁵, with critical regions labelled. Transmembrane helix M2 (green) hinges around the glycine in the centre of the segment, opening the entrance to the inner cavity. The blue volume represents free access of water. The channel is closed (left) at the volume coloured in red, where there is insufficient space for a water molecule to pass. Reproduced, with permission, from ref. 15. **b**, Three subunits of a consensus Kir6.2 model¹⁷ are shown with the rear-most subunit removed for clarity. The subunit positioned to the front provides the bulk of one ATP-binding site in its green carboxy-terminal cytoplasmic domain. The amino-terminal domain of this subunit provides the blue 'slide helix' overlying the binding site to the right and then curves back to close the binding pocket around the adenine ring. Beyond the slide helix, at the turn of the N terminus, residue Arg 50 from the subunit to the left (yellow) is predicted to close the binding pocket at the γ -phosphate of ATP. ATP is rendered by atom type, with oxygens in red and nitrogens in blue.

Kir6.2 constructs have been isolated and studied using electron microscopy and three-dimensional reconstruction³³ (Fig. 1c). The complex is a compact structure of approximately 18 × 13 nm. It points to the intimate packing of four SUR1 subunits against four Kir6.2 subunits, with the TMD0 domain nestled between TMD1, TMD2 and Kir6.2 (Fig. 1c), and suggests that NBF1 and NBF2 interactions can occur between subunits, a possibility that is yet to be experimentally tested.

Mechanism of nucleotide gating of K_{ATP} channels

Several ligands affect K_{ATP} channel activity. ATP (with or without Mg²⁺) inhibits, and PIP₂ activates, by direct interaction with Kir6.2 subunits. Sulphonylureas inhibit, and K⁺ channel-opener drugs activate, by interaction with the SUR subunit. In addition, in the presence of Mg²⁺, ATP and ADP can activate the channel through interaction with the NBFs of SUR. Inhibition by ATP binding to Kir6.2 and activation by Mg-nucleotides is almost certainly the primary physiological regulatory mechanism. Figure 4 provides a theoretical model of the biochemical details. How much material do we have to add realistic details to the model?

Where are the gates?

The weight of evidence suggests that K_{ATP} channels contain two independent types of gate in the permeation pathway (Fig. 4). K_{ATP} chan-

nels show fast, ligand-independent gating (Box 1), which indicates the presence of a gate in the selectivity filter³⁴ (Fig. 2a) — the part of the channel that confers selectivity to particular ions — but here we concern ourselves mainly with the slow, ligand-dependent gates. Crystallization of the bacterial MthK K⁺ channel in an apparently open state led to a now widely accepted hypothesis³⁵ that the main gate in all cation channels involves a hinged motion of M2 (termed S6 in Kv channels; see p. 463), closing the channel at the bottom of the inner cavity, at the narrow collar generated by the crossing of the bundle of M2 helices (Fig. 2a). Consistent with this hypothesis, mutations in M2, particularly near the cytoplasmic end, affect slow gating in multiple Kir channels, including Kir6.2 (refs 36–38). Reduction of single-channel conductance after introduction of positive charge to the inner cavity by methyl thio sulphhydryl ethylamine (MTSEA⁺) modification of introduced cysteines^{39,40} shows that this region can act as a significant barrier in the permeation process. Furthermore, the ability of MTSEA⁺ and other sulphhydryl reagents to access the inner cavity through this region indicates a minimum in the open state diameter. Molecular models of open versus closed states suggest that minimal backbone M2 movement^{15,39,41} is required to increase the diameter of the pore through this region considerably (Fig. 2b).

Kinetic analyses showing that the inner cavity is only accessible to cytoplasmic MTSEA⁺ in the open state and that blockers are trapped in the inner cavity in the closed state^{42,43} seem to confirm that the ligand-operated gate is located at the bottom of the inner cavity (Fig. 2). At this point, there is debate about whether permeation is stopped by specific side chains blocking ion flow, or is a more diffuse reflection of an unfavourable environment for K⁺ dehydration. Whether the final opening transition reflects a concerted motion of all four subunits or whether there are transitional steps involving each subunit is also unresolved^{36,38}, but available experimental data are well described by models in which each subunit undergoes independent gating transitions (Box 1).

How does ATP close the channel?

The cytoplasmic domains of each Kir6.2 subunit provide the binding sites for ATP and PIP₂, which stabilize closed and open states of the channel, respectively³⁶ (Box 1). Molecular modelling suggests that

there are four ATP-binding sites on the channel^{16,17,36}, each formed at the Kir6.2 subunit interfaces. Kinetic analyses indicate that unliganded subunits (that is, those not bound to ATP or PIP₂) are relatively unstable and that ATP and PIP₂ binding are mutually excluded³⁶ (Box 1). Subunits will therefore be predominantly bound to one ligand or the other at any one time. In the absence of ATP, wild-type subunits are bound to membrane PIP₂ (and hence in the permissive, 'open' state) for ~90% of the time. Because all four subunits need to be in the permissive ('open') state for the channel to be active, this results in an observed open probability of ~0.5 (0.9⁴). Binding of ATP to the unliganded subunit will trap it in the non-permissive, 'closed', state. With an association constant, *K*_A, of ~10 μM for wild-type subunits, and with an ATP concentration that is two to three orders of magnitude higher than this in energized cells, each wild-type subunit is probably bound for more than 90% of the time. Trapping of any one subunit in the closed conformation is sufficient to close the channel^{44,45}, and so for most cells under any but the most extreme conditions, the expected open probability would be <0.1⁴; that is, negligible without the stimulatory effects of Mg-nucleotides acting at the SUR subunit (see below).

Docking of ATP in the consensus model of Kir6.2 (ref. 17 and Fig. 2b) was guided by the results of numerous mutagenesis studies that have provided candidate residues for the ATP-binding site. Many mutations alter ATP sensitivity independently of ATP-binding affinity, by altering the intrinsic stability of the open state (that is, they control the open–closed equilibrium constant, *L*; Box 1) and only a few residues (for example, Arg 50 in the N terminus, and Ile 182, Lys 185, Arg 201 and Gly 334 in the C terminus^{46–50}) control ATP affinity directly. These residues are well separated in the primary sequence, but are all close to one another at the inter-subunit interface in the model (Fig. 2b). Potential PIP₂-interacting residues (Arg 54, Arg 176, Arg 177 and Arg 206)^{20,22,49} are near the ATP site and appropriately located to interact with PIP₂ headgroups (Fig. 3a). Overlapping ATP- and PIP₂-binding sites, and biochemical competition of the two ligands for the non-ligand-bound state⁵¹, are consistent with the proposed 'negative heterotropic' interaction of these two ligands in channel regulation^{22,36}.

How the binding of ATP or of PIP₂ is coupled to the closing or opening of the gate is still unknown. One interesting feature of the KirBac structures is the existence of a 'slide helix' (Fig. 2) that lies along the plane of the membrane and physically connects the N terminus, which in Kir6.2 forms part of the ATP-binding site, to the cytoplasmic end of M1 (refs 14, 15). It is interesting that naturally occurring disease mutations that affect the channel open-state stability cluster along the slide helix of Kir6.2 (Fig. 5; see below), highlighting a role for this helix in controlling channel gating^{14,15}.

What goes on at the SUR NBFs?

In the face of ATP inhibition at Kir6.2, physiological activation of the K_{ATP} channel arises acutely from the interaction of nucleotides with the NBFs of SUR⁵² (Fig. 3b). In bacterial NBFs, dimerization is favoured by the presence of nucleotide, and ATP hydrolysis occurs in a cooperative manner^{53,54}, leading to the hypothesis that the nucleotide-bound dimer is the catalytically active species. By analogy, it is suggested that Mg²⁺-dependent ATP hydrolysis at the dimeric SUR NBFs also provides the 'power stroke' that overcomes the inhibitory effect of ATP on Kir6.2 (Fig. 4). The Mg²⁺ dependence of α³²P- and γ³²P-azido-ATP labelling of SUR NBFs provides indirect evidence that SUR NBFs do hydrolyse ATP^{55,56}. Trapping of channels in an activated state by vanadate (a transition-state analogue that mimics a post-hydrolytic, ADP-bound state) and in an inactivated state by beryllium fluoride (a transition-state analogue that mimics the pre-hydrolytic, ATP-bound state)⁵⁷ add further consistency to the hypothesis that ATP hydrolysis underlies channel activation.

However, direct hydrolysis measurements on SUR NBFs are sparse^{33,57–59}, and mutations analogous to those that have drastic effects on ATP hydrolysis in bacterial NBFs have only modest effects on SUR NBFs⁵⁸. In patch-clamp experiments, channel activation by exogenous MgADP is much more effective than activation by MgATP, and the

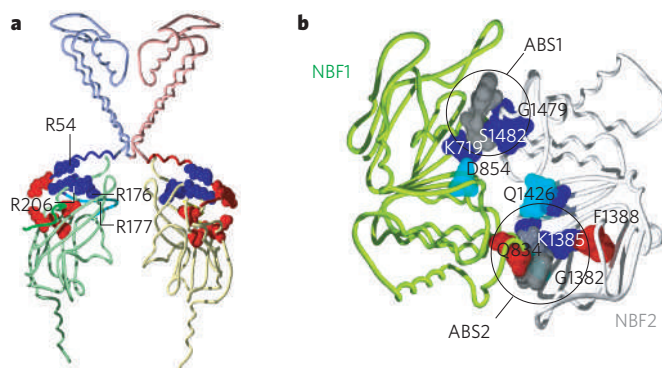


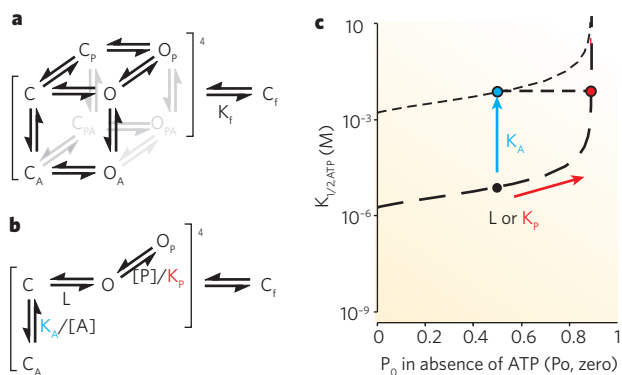
Figure 3 | Binding sites for opening ligands. **a**, Molecular basis of the 'negative heterotropic' interaction between ATP and PIP₂. Predicted ATP-binding site residues (red) and potential PIP₂-interacting residues (blue) in a consensus Kir6.2 model⁹² show how close these groups are likely to be, and how potential PIP₂-binding residues are in the region of phospholipid headgroups. Two diagonally opposed subunits are shown in each case. **b**, Nucleotide-binding fold 1 (NBF1)–NBF2 heterodimeric structure of ATP-binding sites (ABSs) in sulphonylurea receptor SUR1 (modelled on the bacterial ATP-binding protein MsA using MODELER version 6, verified with PROCHECK). ATP is rendered in grey in ABS1 and ABS2. Mutation of residues in dark and light blue causes reduced or slowed MgADP activation of channel activity, respectively. Mutation of residues in red causes enhanced MgADP activation.

Box 1 | Kinetic mechanism of K_{ATP} channel gating

Even in the absence of ATP, the single-channel kinetics of K_{ATP} are complex, but certain key features are clearly discernible and can be modelled by assuming a single 'fast' gate and ligand-operated 'slow' gates within each subunit³⁶. Single-channel analyses consistently reveal one open time and multiple closed times^{34,44,50}. The fast gate gives rise to the short open and closed times (the 'intra-burst' events), which are voltage-dependent but ligand-independent, and affected by mutations of residues in or near the selectivity filter³⁴. In addition, there are several longer closed times (the 'inter-burst' closures). It is these slow gating events that are modulated by nucleotides and membrane phosphatidylinositol-4,5-bisphosphate (PIP_2)^{38,44}.

Tetrameric models (shown in panel **a**), in which each subunit undergoes independent gating transitions, not only replicate all essential features of ATP inhibition, but also have important predictive properties³⁶. The assumption that each of the four subunits can be in an open or closed conformation (in the absence of ligand) and that the channel conducts only if all subunits are in the open conformation automatically produces multiple closed states⁴⁴ and implies that closure of any one subunit will be sufficient to close the channel⁴⁵. Assuming that the open channel can close by such individual subunit closures (O_P to C_P or O_A to C_A) or, at the selectivity filter, by a concerted event (fast gating, O to C_f controlled by K_f), such a model produces a single short open time and multiple closed times with inter-burst closures dominated by subunit closures³⁶. The physical reality is that PIP_2 - and ATP-binding sites are probably overlapping (Fig. 3a), consistent with negligible occupancy of C_{PA} and O_{PA} states predicted by the model. Furthermore, O_A and C_P states are rarely sampled in wild-type models.

For wild-type channels, the subunits then exist primarily in only four states (panel **b**) and ATP binds primarily to the closed state. This feature predicts a characteristic nonlinear relationship (panel **c**) between ATP sensitivity (the concentration of ATP causing half-maximal inhibition, $K_{1/2,ATP}$) and open probability (P_O) — one that is clearly adhered to by wild-type and mutant channels under a range of experimental conditions³⁶. Changes in intrinsic open-state stability (L) or PIP_2 affinity (K_p) cause the channel to move up or down this nonlinear relationship, whereas changes in ATP affinity (K_A) shift the channel onto a parallel relationship. One consequence of these two effects is that it is possible to reduce apparent ATP sensitivity (as measured by $K_{1/2,ATP}$) to the same value by two mechanistic routes (dashed line), with different consequences for other regulatory features (see main text).



same 'hydrolysis' mutations lead to decreased MgADP stimulation of K_{ATP} channels^{58,60–62} (Fig. 3b). Why MgADP should so strongly activate the channel in excised patches if a hydrolytic cycle is required for activation is unclear. To fit this property into the model (Fig. 4) would require that the 'activated' state resulting from hydrolysis persists after MgADP dissociation, at least long enough for rebinding of exogenous MgADP to maintain the state. Just how reasonable this notion is has not been addressed. It remains conceivable that hydrolysis, if it happens at all, is an epiphenomenon and that preferential binding of MgADP at ABS2 is the activator of the channel.

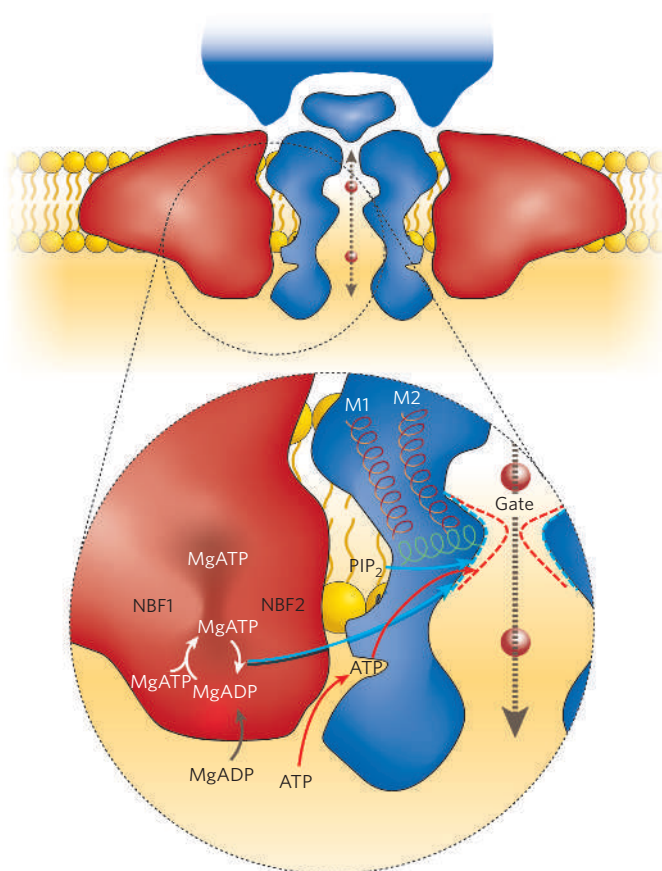


Figure 4 | Nucleotide regulation of K_{ATP} activity. The metabolically controlled gate of the channel is located at the cytoplasmic end of the inner cavity. Phosphatidylinositol-4,5-bisphosphate (PIP_2) interaction provides an energetic pull to open channels, and ATP binding provides the energetic push to close the ligand-operated gate, perhaps through the physical link provided by the 'slide helix' (green). MgATP binds to each of the ATP-binding sites (ABSs) that are formed at the interface between nucleotide-binding folds NBF1 and NBF2 in each sulphonylurea receptor (SUR) subunit. ATP hydrolysis at the second site results in a conformational 'activated' state that is transduced to an 'override' of ATP inhibition at the Kir6.2 subunit. The 'activated' state persists through MgADP dissociation, and can be maintained by MgADP rebinding.

From structure to disease and back

K_{ATP} channels are widely expressed in vertebrate tissues, and genetic manipulation in mice has revealed important consequences of loss of channel subunits in different tissues, highlighting the coordinate nature of the two subunits in generating tissue-specific channel activity⁶³. The same tissue-specific phenotype is generated in several cases when one or the other partner is knocked out: both $SUR1^{-/-}$ and $Kir6.2^{-/-}$ reiterate a glucose-insensitive insulin secretory phenotype due to loss of K_{ATP} in the pancreas^{64–66}. $SUR2^{-/-}$ and $Kir6.2^{-/-}$ animals both lack muscle K_{ATP} ^{67,68}, and both $SUR2^{-/-}$ and $Kir6.1^{-/-}$ reiterate a phenotype that mimics human Prinzmetal angina^{69,70}, with spontaneous coronary vasospasm leading to early death, probably due to the vasoconstrictive effects of reduced vascular K_{ATP} current. Overexpression of mutant K_{ATP} subunits in the pancreas also shows how increased or decreased channel activity, respectively, underlies a hypoinsulinaemic, diabetic phenotype, or hyperinsulinaemic phenotype, respectively⁷¹ (Fig. 5a). Alongside these animal models, genetic analyses have demonstrated corollary diseases—neonatal diabetes mellitus (NDM) and hyperinsulinaemia—in humans^{1–3,71,72}. Studies of recombinant properties of mutant channels can explain the molecular basis of such diseases, and can also provide important forward genetic evidence for the role of specific structural features of the channel.

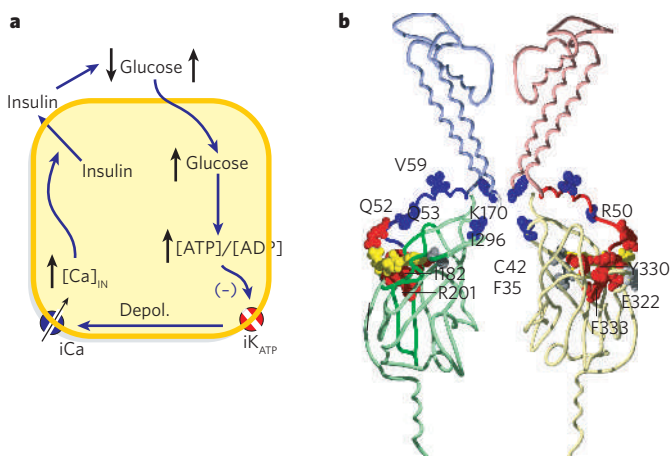


Figure 5 | Molecular defects of neonatal diabetes. **a**, When glucose rises in the pancreatic β -cell, metabolism is stimulated, raising the ratio of ATP to ADP. This closes K_{ATP} channels, which depolarizes the cell, opening Ca^{2+} channels, causing intracellular Ca^{2+} to rise, and triggering insulin secretion. Insensitivity of K_{ATP} to ATP inhibition will cause maintained hyperpolarization and decreased insulin secretion. **b**, Residues of inward rectifier K^+ channel subunit Kir6.2 that are mutated in neonatal diabetes, highlighted in the consensus model⁹². Residues in red contribute to forming the ATP-binding site. Residues in blue, conspicuously clustering at the end of the inner cavity and in the 'slide helix', control intrinsic open-state stability. Residues in grey have not yet been studied in detail. ATP is shown in yellow. Two opposing subunits are shown (subunit *a*, *c* in the transmembrane region; *b*, *d* in the cytoplasmic region).

K_{ATP} defects in hyperinsulinaemia

K_{ATP} channels suppress insulin secretion by hyperpolarizing the β -cell membrane (Fig. 5a), and so loss of K_{ATP} channel activity is expected to cause hypersecretion. Congenital hyperinsulinaemia, a genetic disorder characterized by dysregulated insulin secretion, is the most common cause of hypoglycaemia in infancy, and more than 50 loss-of-function hyperinsulinaemia mutations have now been recognized in K_{ATP} genes⁷³. Although some mutations are in Kir6.2, the majority are in SUR1 (refs 74,75) and most cause one of two major defects: biosynthetic or trafficking defects that lead to absent or reduced surface expression of channels; or loss of MgADP activation of expressed channels, which could reflect direct effects on ATP hydrolysis or ADP binding at the NBFs, or defects in the coupling to channel opening. As a result, β -cells remain depolarized and persistently secreting. Hyperinsulinaemia mutations tend to cluster in the SUR1 NBFs, and almost invariably result in reduced MgADP activation⁷⁶ (Fig. 3b), highlighting the crucial role of SUR NBFs in the physiological activation of K_{ATP} channels.

One of the most common hyperinsulinaemia mutations (SUR1 [Δ Phe 1388]) causes defective trafficking and lack of surface expression of functional K_{ATP} channels, and altered channel function⁷⁷. The mutant protein appears to be retained in the ER, similar to the CFTR [Δ Phe 508] mutation (see p. 477). Another single amino acid mutation (Arg1394His) also causes defective trafficking in mammalian cells, but rather than being retained in the ER, the protein appears to accumulate in the Golgi⁷⁸. For such mutations, manipulations that allow correction of the trafficking defects might be of therapeutic value⁷⁹.

K_{ATP} defects in neonatal diabetes

Whereas loss of K_{ATP} activity causes persistent hypersecretion of insulin, gain of function should cause the opposite: that is, lack of secretion and diabetes. Multiple mutations in Kir6.2 have recently been found to underlie NDM. Random mutations are most likely to cause loss of a specific protein function. In this case, it is loss of ATP sensitivity^{71,72,80} resulting from one of two mechanisms (Box 1): decreased apparent affinity of the ATP-binding pocket for the nucleotide; or an increase in the intrinsic stability of the open state^{81,82}. Consistent with a direct

effect on binding, residues Arg 50, Ile 182 and Arg 201, all identified as mutated in NDM patients, were all previously implicated as likely ATP-binding residues in Kir6.2, and all are located in the ATP-binding site in the model channel (Fig. 5b). Tyr 330 and Phe 333, also found to be mutated in NDM patients, are predicted to lie close to the phosphate tail in the binding pocket¹⁶.

Mutations that alter channel gating in the absence of ATP can occur throughout the Kir6.2 subunit^{44,50,83}. By stabilizing the open state of the channel (increasing *L*; Box 1), ATP sensitivity is reduced without a change in ATP-binding affinity^{36,84}. Loss of ATP sensitivity coupled to high open probability in Gln52Arg, Cys42Arg, Tyr330Cys, Ile1296Leu and Val59Gly mutations indicates that this mechanism is operating in these NDM mutations^{81,82,85}. Gln 52 and Val 59 are located within the slide helix region of Kir6.2 (Fig. 6), lending consistency to the suggestion that this 'linker' serves to physically couple the ATP-binding site to the gating region¹⁴. An effect on the gate region itself might explain the diabetes-causing effects of mutations at residue Lys 170 (Lys170Asn, Lys170Arg)⁸⁶ (Fig. 5b).

The realization that neonatal diabetes results from gain of K_{ATP} channel function has rapidly led to changes in therapy from injected insulin to the painless use of oral sulphonylureas^{87,88}. However, understanding of the molecular basis of K_{ATP} function also immediately predicts an underlying caveat: sulphonylureas also stabilize the closed state of the channel, and sulphonylurea sensitivity depends on the open-state stability of the channel, such that mutations that increase open-state stability also reduce sulphonylurea sensitivity⁸². Because different NDM mutations alter ATP sensitivity by two distinct mechanisms (by reducing ATP affinity or by increasing the open-state stability; Box 1), so they show differential sulphonylurea sensitivity⁸²—mechanistic information that might help in tailoring pharmacologies for treating the disease.

Future directions

What are the important questions to pursue in future? From a strictly reductionist angle, the first challenge is to generate a high-resolution crystal structure of the K_{ATP} complex. The recent structural study by Mikhailov *et al.*³³ is tantalizing, suggesting that recombinant expression in insect cells will prove to be a useful approach. Beyond this, hints that K_{ATP} can be complexed with enzymes and other proteins^{89,90} need to be explored using cellular imaging approaches, as well as with recombinant complexes that could lead to higher-resolution structures. Less ambitiously, short-term gains will be made from crystallizing domains of the channel. Crystallization of the cytoplasmic domains of Kir2.1 (ref. 13) and Kir3.1 (ref. 12), as well as the NBFs of CFTR (see p. 477), indicates that crystallization of the homologous ligand-interacting domains in the K_{ATP} channel could be realized in the near future. Further progress in crystallizing KirBac proteins, combined with biochemical and functional assays, will provide increasing insights into gating mechanisms and the molecular details of channel–ligand interactions.

What are the frontiers in terms of understanding ligand regulation and gating motions in the channel? There is now a consensus that ligand-dependent gating involves motions of M2/S6 that block permeation at the bottom of the inner cavity, but exactly how this is coupled to ligand interactions with cytoplasmic domains is unclear. The conformational changes that result from nucleotide hydrolysis and ADP binding at the NBFs of SUR, or the interaction of sulphonylureas and K^+ channel openers with SUR, and how they ultimately influence the channel gate, are also unknown.

How will future structural insights affect our understanding of pathophysiology or our approach to clinical therapeutics? K_{ATP} could be involved in a whole spectrum of diabetes mellitus⁸⁵ and other organ pathologies⁶³, and we will continue to see effects of K_{ATP} subunit deletion and manipulation affecting the functioning of multiple organs. Because all proteins are evolutionarily tailored to perform specific functions, it follows that disease-causing mutations are most likely to cause loss of the specific function. In the present context, loss of channel activation due to loss of the activating effects of nucleotides on SUR1 underlies net

channel 'loss of function' in most cases of hyperinsulinaemia, whereas loss of ATP sensitivity through mutation of Kir6.2 underlies net 'gain of function' in all NDM mutations identified so far. But it is conceivable that spontaneous mutations also cause increases in an activating function, such as increased MgADP affinity of SUR1 (refs 60, 91). Detection of disease-causing mutations will further inform the molecular basis of channel function, but the sulphonylurea desensitizing effect of open-state stabilizing mutations⁸² shows how an understanding of the underlying molecular basis of activity can provide insights into the mechanisms of disease. ■

- Miki, T. & Seino, S. Roles of K_{ATP} channels as metabolic sensors in acute metabolic changes. *J. Mol. Cell. Cardiol.* **38**, 917–925 (2005).
- Bryan, J., Vila-Carriels, W. H., Zhao, G., Babenko, A. P. & Aguilar-Bryan, L. Toward linking structure with function in ATP-sensitive K^+ channels. *Diabetes* **53** (Suppl. 3), S104–S112 (2004).
- Ashcroft, F. M. ATP-sensitive potassium channelopathies: focus on insulin secretion. *J. Clin. Invest.* **115**, 2047–2058 (2005).
- Flagg, T. P. & Nichols, C. Sarcolemmal K_{ATP} channels: what do we really know? *J. Mol. Cell. Cardiol.* **39**, 61–70 (2005).
- Inagaki, N. *et al.* Reconstitution of I_{KATP} : an inward rectifier subunit plus the sulphonylurea receptor. *Science* **270**, 1166–1170 (1995).
- Inagaki, N. *et al.* A family of sulphonylurea receptors determines the pharmacological properties of ATP-sensitive K^+ channels. *Neuron* **16**, 1011–1017 (1996).
- Yamada, M. *et al.* Sulphonylurea receptor 2B and Kir6.1 form a sulphonylurea-sensitive but ATP-insensitive K^+ channel. *J. Physiol. (Lond.)* **499**, 715–720 (1997).
- Shyng, S. & Nichols, C. G. Octameric stoichiometry of the K_{ATP} channel complex. *J. Gen. Physiol.* **110**, 655–664 (1997).
- Clement, J. P. IV *et al.* Association and stoichiometry of K_{ATP} channel subunits. *Neuron* **18**, 827–838 (1997).
- Neagoe, I. & Schwappach, B. Pas de deux in groups of four—the biogenesis of K_{ATP} channels. *J. Mol. Cell. Cardiol.* **38**, 887–894 (2005).
- Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* **280**, 69–77 (1998).
- Nishida, M. & MacKinnon, R. Structural basis of inward rectification. Cytoplasmic pore of the G protein-gated inward rectifier GIRK1 at 1.8 Å resolution. *Cell* **111**, 957–965 (2002).
- Pegan, S. *et al.* Cytoplasmic domain structures of Kir2.1 and Kir3.1 show sites for modulating gating and rectification. *Nature Neurosci.* **8**, 279–287 (2005).
- Kuo, A. *et al.* Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science* **300**, 1922–1926 (2003).
- Kuo, A., Domene, C., Johnson, L. N., Doyle, D. A. & Venien-Bryan, C. Two different conformational states of the KirBac3.1 potassium channel revealed by electron crystallography. *Structure* **13**, 1463–1472 (2005).
- Antcliff, J. F., Haider, S., Proks, P., Sansom, M. S. & Ashcroft, F. M. Functional analysis of a structural model of the ATP-binding site of the K_{ATP} channel Kir6.2 subunit. *EMBO J.* **24**, 229–239 (2005).
- Haider, S., Antcliff, J. F., Proks, P., Sansom, M. S. & Ashcroft, F. M. Focus on Kir6.2: a key component of the ATP-sensitive potassium channel. *J. Mol. Cell. Cardiol.* **38**, 927–936 (2005).
- Enkvetchakul, D. *et al.* Functional characterization of a prokaryotic Kir channel. *J. Biol. Chem.* **279**, 47076–47080 (2004).
- Enkvetchakul, D., Jeliakova, I. & Nichols, C. G. Direct modulation of Kir channel gating by membrane phosphatidylinositol 4,5-bisphosphate. *J. Biol. Chem.* **280**, 35785–35788 (2005).
- Baukowitz, T. *et al.* PIP₂ and PIP as determinants for ATP inhibition of K_{ATP} channels. *Science* **282**, 1141–1144 (1998).
- Fan, Z. & Makielski, J. C. Anionic phospholipids activate ATP-sensitive potassium channels. *J. Biol. Chem.* **272**, 5388–5395 (1997).
- Shyng, S. L. & Nichols, C. G. Membrane phospholipid control of nucleotide sensitivity of K_{ATP} channels. *Science* **282**, 1138–1141 (1998).
- Chan, K. W., Zhang, H., Mishahi, T. & Logothetis, D. E. Characterization of the interaction between the N-terminal transmembrane domain of the sulphonylurea receptor (SUR1) and Kir6.2. *Biophys. J.* **82**, 590a (2002).
- Babenko, A. P. & Bryan, J. SUR-dependent modulation of K_{ATP} channels by an N-terminal Kir6.2 peptide. Defining intersubunit gating interactions. *J. Biol. Chem.* **277**, 43997–44004 (2002).
- Babenko, A. P. & Bryan, J. SUR domains that associate with and gate K_{ATP} pores define a novel gatekeeper. *J. Biol. Chem.* **278**, 41577–41580 (2003).
- Campbell, J. D., Proks, P., Lippat, J. D., Sansom, M. S. & Ashcroft, F. M. Identification of a functionally important negatively charged residue within the second catalytic site of the SUR1 nucleotide-binding domains. *Diabetes* **53** (Suppl. 3), S123–S127 (2004).
- Hung, L. W. *et al.* Crystal structure of the ATP-binding subunit of an ABC transporter. *Nature* **396**, 703–707 (1998).
- Hopfner, K. P. *et al.* Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily. *Cell* **101**, 789–800 (2000).
- Locher, K. P., Lee, A. T. & Rees, D. C. The *E. coli* BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science* **296**, 1091–1098 (2002).
- Matsuo, M., Kioka, N., Amachi, T. & Ueda, K. ATP binding properties of the nucleotide-binding folds of SUR1. *J. Biol. Chem.* **274**, 37479–37482 (1999).
- Hough, E., Mair, L., Mackenzie, W. & Sivaprasadarao, A. Expression, purification, and evidence for the interaction of the two nucleotide-binding folds of the sulphonylurea receptor. *Biochem. Biophys. Res. Commun.* **294**, 191–197 (2002).
- Mikhailov, M. V. & Ashcroft, S. J. H. Interactions of the sulphonylurea receptor 1 subunit in the molecular assembly of β -cell K-ATP channels. *J. Biol. Chem.* **275**, 3360–3364 (2000).
- Mikhailov, M. V. *et al.* 3-D structural and functional characterization of the purified K_{ATP} channel complex Kir6.2–SUR1. *EMBO J.* **24**, 4166–4175 (2005).
- Proks, P., Capener, C. E., Jones, P. & Ashcroft, F. M. Mutations within the P-loop of Kir6.2 modulate the intraburst kinetics of the ATP-sensitive potassium channel. *J. Gen. Physiol.* **118**, 341–353 (2001).
- Jiang, Y. *et al.* The open pore conformation of potassium channels. *Nature* **417**, 523–526 (2002).
- Enkvetchakul, D. & Nichols, C. G. Gating mechanism of K_{ATP} channels: function fits form. *J. Gen. Physiol.* **122**, 471–480 (2003).
- Sackin, H., Nanazashvili, M., Palmer, L. G., Krambis, M. & Walters, D. E. Structural locus of the pH gate in the Kir1.1 inward rectifier channel. *Biophys. J.* **88**, 2597–2606 (2005).
- Drain, P., Geng, X. & Li, L. Concerted gating mechanism underlying K_{ATP} channel inhibition by ATP. *Biophys. J.* **86**, 2101–2112 (2004).
- Loussouarn, G., Phillips, L. R., Masia, R., Rose, T. & Nichols, C. G. Flexibility of the Kir6.2 inward rectifier K^+ channel pore. *Proc. Natl Acad. Sci. USA* **98**, 4227–4232 (2001).
- Lu, T., Nguyen, B., Zhang, X. & Yang, J. Architecture of a K^+ channel inner pore revealed by stoichiometric covalent modification. *Neuron* **37**, 571–580 (1999).
- Domene, C., Doyle, D. A. & Venien-Bryan, C. Modeling of an ion channel in its open conformation. *Biophys. J.* **89**, L01–L03 (2005).
- Phillips, L. R. & Nichols, C. G. Ligand-induced closure of inward rectifier Kir6.2 channels traps spermine in the pore. *J. Gen. Physiol.* **122**, 795–804 (2003).
- Phillips, L. R., Enkvetchakul, D. & Nichols, C. G. Gating dependence of inner pore access in inward rectifier K^+ channels. *Neuron* **37**, 953–962 (2003).
- Enkvetchakul, D., Loussouarn, G., Makhina, E., Shyng, S. L. & Nichols, C. G. The kinetic and physical basis of K_{ATP} channel gating: toward a unified molecular understanding. *Biophys. J.* **78**, 2334–2348 (2000).
- Markworth, E., Schwanstecher, C. & Schwanstecher, M. ATP⁴⁺ mediates closure of pancreatic β -cell ATP-sensitive potassium channels by interaction with 1 of 4 identical sites. *Diabetes* **49**, 1413–1418 (2000).
- Proks, P., Gribble, F. M., Adhikari, R., Tucker, S. J. & Ashcroft, F. M. Involvement of the N-terminus of Kir6.2 in the inhibition of the K_{ATP} channel by ATP. *J. Physiol. (Lond.)* **514**, 19–25 (1999).
- Li, L., Wang, J. & Drain, P. The I182 region of Kir6.2 is closely associated with ligand binding in K_{ATP} channel inhibition by ATP. *Biophys. J.* **79**, 841–852 (2000).
- Tucker, S. J., Gribble, F. M., Zhao, C., Trapp, S. & Ashcroft, F. M. Truncation of Kir6.2 produces ATP-sensitive K^+ channels in the absence of the sulphonylurea receptor. *Nature* **387**, 179–183 (1997).
- Shyng, S. L., Cukras, C. A., Harwood, J. & Nichols, C. G. Structural determinants of PIP₂ regulation of inward rectifier K_{ATP} channels. *J. Gen. Physiol.* **116**, 599–608 (2000).
- Drain, P., Li, L. & Wang, J. K_{ATP} channel inhibition by ATP requires distinct functional domains of the cytoplasmic C terminus of the pore-forming subunit. *Proc. Natl Acad. Sci. USA* **95**, 13953–13958 (1998).
- MacGregor, G. G. *et al.* Nucleotides and phospholipids compete for binding to the C terminus of K_{ATP} channels. *Proc. Natl Acad. Sci. USA* **99**, 2726–2731 (2002).
- Nichols, C. G. *et al.* Adenosine diphosphate as an intracellular regulator of insulin secretion. *Science* **272**, 1785–1787 (1996).
- Lammens, A., Schele, A. & Hopfner, K. P. Structural biochemistry of ATP-driven dimerization and DNA-stimulated activation of SMC ATPases. *Curr. Biol.* **14**, 1778–1782 (2004).
- Smith, P. C. *et al.* ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. *Mol. Cell* **10**, 139–149 (2002).
- Matsuo, M., Tanabe, K., Kioka, N., Amachi, T. & Ueda, K. Different binding properties and affinities for ATP and ADP among sulphonylurea receptor subtypes, SUR1, SUR2A, and SUR2B. *J. Biol. Chem.* **275**, 28757–28763 (2000).
- Ueda, K., Komine, J., Matsuo, M., Seino, S. & Amachi, T. Cooperative binding of ATP and MgADP in the sulphonylurea receptor is modulated by glibenclamide. *Proc. Natl Acad. Sci. USA* **96**, 1268–1272 (1999).
- Zingman, L. V. *et al.* Signaling in channel/enzyme multimers: ATPase transitions in SUR module gate ATP-sensitive K^+ conductance. *Neuron* **31**, 233–245 (2001).
- Bienengraeber, M. *et al.* ATPase activity of the sulphonylurea receptor: a catalytic function for the K_{ATP} channel complex. *FASEB J.* **14**, 1943–1952 (2000).
- Masia, R., Enkvetchakul, D. & Nichols, C. Differential nucleotide regulation of K_{ATP} channels by SUR1 and SUR2A. *J. Mol. Cell. Cardiol.* **39**, 491–501 (2005).
- Shyng, S., Ferrigni, T. & Nichols, C. G. Regulation of K_{ATP} channel activity by diazoxide and MgADP. Distinct functions of the two nucleotide binding folds of the sulphonylurea receptor. *J. Gen. Physiol.* **110**, 643–654 (1997).
- Gribble, F. M., Tucker, S. J. & Ashcroft, F. M. The essential role of the Walker A motifs of SUR1 in K-ATP channel activation by Mg-ADP and diazoxide. *EMBO J.* **16**, 1145–1152 (1997).
- Zingman, L. V. *et al.* Tandem function of nucleotide binding domains confers competence to sulphonylurea receptor in gating ATP-sensitive K^+ channels. *J. Biol. Chem.* **277**, 14206–14210 (2002).
- Seino, S. & Miki, T. Gene targeting approach to clarification of ion channel function: studies of Kir6.x null mice. *J. Physiol. (Lond.)* **554**, 295–300 (2004).
- Miki, T. *et al.* Defective insulin secretion and enhanced insulin action in K_{ATP} channel-deficient mice. *Proc. Natl Acad. Sci. USA* **95**, 10402–10406 (1998).
- Seghers, V., Nakazaki, M., DeMayo, F., Aguilar-Bryan, L. & Bryan, J. Sur1 knockout mice. A model for K_{ATP} channel-independent regulation of insulin secretion. *J. Biol. Chem.* **275**, 9270–9277 (2000).
- Shiota, C. *et al.* Sulphonylurea receptor type 1 knock-out mice have intact feeding-stimulated insulin secretion despite marked impairment in their response to glucose. *J. Biol. Chem.* **277**, 37176–37183 (2002).
- Chutkow, W. A. *et al.* Disruption of Sur2-containing K_{ATP} channels enhances insulin-stimulated glucose uptake in skeletal muscle. *Proc. Natl Acad. Sci. USA* **98**, 11760–11764 (2001).
- Suzuki, M. *et al.* Functional roles of cardiac and vascular ATP-sensitive potassium channels clarified by Kir6.2-knockout mice. *Circ. Res.* **88**, 570–577 (2001).
- Miki, T. *et al.* Mouse model of Prinzmetal angina by disruption of the inward rectifier Kir6.1.

- Nature Med.* **8**, 466–472 (2002).
70. Chutkow, W. A. *et al.* Episodic coronary artery vasospasm and hypertension develop in the absence of Sur2 K_{ATP} channels. *J. Clin. Invest.* **110**, 203–208 (2002).
 71. Koster, J. C., Permutt, A. & Nichols, C. G. Diabetes and insulin secretion: the ATP-sensitive K^+ channel (K_{ATP}) connection. *Diabetes* **54**, 3065–3072 (2005).
 72. Hattersley, A. T. & Ashcroft, F. M. Activating mutations in Kir6.2 and neonatal diabetes: new clinical syndromes, new scientific insights, and new therapy. *Diabetes* **54**, 2503–2513 (2005).
 73. Sharma, N., Crane, A., Gonzalez, G., Bryan, J. & Aguilar-Bryan, L. Familial hyperinsulinism and pancreatic β -cell ATP-sensitive potassium channels. *Kidney Int.* **57**, 803–808 (2000).
 74. Nestorowicz, A. *et al.* Genetic heterogeneity in familial hyperinsulinism. *Hum. Mol. Genet.* **7**, 1119–1128 (1998).
 75. Tornovsky, S. *et al.* Hyperinsulinism of infancy: novel ABCC8 and KCNJ11 mutations and evidence for additional locus heterogeneity. *J. Clin. Endocrinol. Metab.* **89**, 6224–6234 (2004).
 76. Shyng, S. L. *et al.* Functional analyses of novel mutations in the sulfonylurea receptor 1 associated with persistent hyperinsulinemic hypoglycemia of infancy. *Diabetes* **47**, 1145–1151 (1998).
 77. Cartier, E., Conti, L. R., Vandenberg, C. A. & Shyng, S.-L. Defective trafficking and function of K_{ATP} channels caused by a sulfonylurea receptor 1 mutation associated with persistent hyperinsulinemic hypoglycemia of infancy. *Proc. Natl Acad. Sci. USA* **98**, 2882–2887 (2001).
 78. Partridge, C. J., Beech, D. J. & Sivaprasadarao, A. Identification and pharmacological correction of a membrane trafficking defect associated with a mutation in the sulfonylurea receptor causing familial hyperinsulinism. *J. Biol. Chem.* **276**, 35947–35952 (2001).
 79. Yan, F. *et al.* Sulfonylureas correct trafficking defects of ATP-sensitive potassium channels caused by mutations in the sulfonylurea receptor. *J. Biol. Chem.* **279**, 11096–11105 (2004).
 80. Gloyn, A. L. *et al.* Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N. Engl. J. Med.* **350**, 1838–1849 (2004).
 81. Proks, P. *et al.* Molecular basis of Kir6.2 mutations associated with neonatal diabetes plus neurological features. *Proc. Natl Acad. Sci. USA* **101**, 17539–17544 (2004).
 82. Koster, J. C., Remedi, M. S., Dao, C. & Nichols, C. G. ATP and sulfonylurea sensitivity of mutant ATP-sensitive K^+ channels in neonatal diabetes: implications for pharmacogenomic therapy. *Diabetes* **54**, 2645–2654 (2005).
 83. Tucker, S. J. *et al.* Molecular determinants of K_{ATP} channel inhibition by ATP. *EMBO J.* **17**, 3290–3296 (1998).
 84. Schwanstecher, C., Meyer, U. & Schwanstecher, M. $K_{IR6.2}$ polymorphism predisposes to type 2 diabetes by inducing overactivity of pancreatic β -cell ATP-sensitive K^+ channels. *Diabetes* **51**, 875–879 (2002).
 85. Yorifuji, T. *et al.* The C42R mutation in the Kir6.2 (KCNJ11) gene as a cause of transient neonatal diabetes, childhood diabetes, or later-onset, apparently type 2 diabetes mellitus. *J. Clin. Endocrinol. Metab.* **90**, 3174–3178 (2005).
 86. Massa, O. *et al.* KCNJ11 activating mutations in Italian patients with permanent neonatal diabetes. *Hum. Mutat.* **25**, 22–27 (2005).
 87. Sagen, J. V. *et al.* Permanent neonatal diabetes due to mutations in KCNJ11 encoding Kir6.2: patient characteristics and initial response to sulfonylurea therapy. *Diabetes* **53**, 2713–2718 (2004).
 88. Zung, A., Glaser, B., Nimri, R. & Zadik, Z. Glibenclamide treatment in permanent neonatal diabetes mellitus due to an activating mutation in Kir6.2. *J. Clin. Endocrinol. Metab.* **89**, 5504–5507 (2004).
 89. Crawford, R. M. *et al.* M-LDH serves as a sarcolemmal K_{ATP} channel subunit essential for cell protection against ischemia. *EMBO J.* **21**, 3936–3948 (2002).
 90. Chowdhury, P. D. *et al.* The glycolytic enzymes, glyceraldehyde-3-phosphate dehydrogenase, triose-phosphate isomerase, and pyruvate kinase are components of the K_{ATP} channel macromolecular complex and regulate its function. *J. Biol. Chem.* **280**, 38464–38470 (2005).
 91. Cartier, E. A., Shen, S. & Shyng, S. L. Modulation of the trafficking efficiency and functional properties of ATP-sensitive potassium channels through a single amino acid in the sulfonylurea receptor. *J. Biol. Chem.* **278**, 7081–7090 (2003).
 92. Haider, S., Grottesi, A., Hall, B. A., Ashcroft, F. M. & Sansom, M. S. Conformational dynamics of the ligand-binding domain of inward rectifier K channels as revealed by molecular dynamics simulations: toward an understanding of Kir channel gating. *Biophys. J.* **88**, 3310–3320 (2005).

Acknowledgements I am grateful to R. Masia, B. Koster, D. Enkvetchakul, H. Kurata and T. Flagg for providing suggestions on the text, and to R. Masia, D. Enkvetchakul, F. Ashcroft and M. Sansom for providing molecular models of SUR1 NBFs and Kir6.2.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to C.G.N. (cnichols@cellbio.wustl.edu).

The ABC protein turned chloride channel whose failure causes cystic fibrosis

David C. Gadsby¹, Paola Vergani² & László Csanády³

CFTR chloride channels are encoded by the gene mutated in patients with cystic fibrosis. These channels belong to the superfamily of ABC transporter ATPases. ATP-driven conformational changes, which in other ABC proteins fuel uphill substrate transport across cellular membranes, in CFTR open and close a gate to allow transmembrane flow of anions down their electrochemical gradient. New structural and biochemical information from prokaryotic ABC proteins and functional information from CFTR channels has led to a unifying mechanism explaining those ATP-driven conformational changes.

The all-too-common deadly disease of young people, cystic fibrosis (CF), is caused by failure of a chloride-ion channel with the unwieldy name of CFTR, standing for cystic fibrosis transmembrane conductance regulator. Anion flow through this channel is needed for normal function of epithelia such as those that line airways and the intestinal tract and ducts in the pancreas, testes and sweat glands. Without anion flow, water movement slows, and dehydrated mucus clogs ducts and collects in the lung where it fosters ultimately lethal bacterial infections. The amino-acid sequence of CFTR immediately identified it as a member of the superfamily of ATP-binding cassette (ABC) transporter ATPases. Among the thousands of ABC family members, only CFTR is an ion channel. High-resolution electrophysiological studies of single CFTR channels therefore offer a unique window into mechanisms of ABC transporter function. And recent structural and biochemical advances in the study of bacterial ABC proteins have shed light on how opening and closing of the CFTR chloride channel are regulated.

Here we bring together the latest information on CFTR-channel function and on prokaryotic ABC-transporter structure to provide an up-to-the-minute view of what CFTR channels normally do, what a CFTR channel might look like and how its anion pathway is opened and closed. The gate that regulates chloride-ion flow through the membrane-spanning pore in CFTR is thought to be opened when ATP binds to CFTR's two cytoplasmic nucleotide-binding domains (NBD1 and NBD2) and, acting as molecular glue, holds them together. Hydrolysis of one of the ATPs disrupts the NBD1–NBD2 interaction and closes the channel gate, terminating anion flow. Kinase-mediated phosphorylation of a cytoplasmic regulatory domain in CFTR is needed for successful transmission of NBD events to the channel gate.

Evolutionary conservation of structural components of the NBD1–NBD2 interface makes it probable that a similar cycle of ATP binding and hydrolysis drives function in most ABC family members. Even so, important differences between NBD1 and NBD2 in CFTR seem to have evolved to allow only one of the bound ATPs to be hydrolysed while the other remains unaltered during numerous gating cycles. Similar differences in the NBDs are found in CFTR's closest relatives in the human genome. These are the multidrug-resistance-related proteins (MRPs), which are ATP-driven efflux pumps, and the sulphonylurea receptors (SURs), which associate with inward rectifier potassium-ion channels to form nucleotide-sensitive potassium conductance (see the review

in this issue by Nichols, p. 470). So mechanisms inferred for CFTR are likely to apply also to MRP and SUR.

Cystic fibrosis and CFTR

In 1989, when the gene mutated in CF patients was identified by positional cloning^{1,2}, the expectation was that it would encode a chloride-ion channel. This was because CF epithelia behaved as though they were impermeable to chloride³, and chloride channels could be activated by cAMP-dependent protein kinase (PKA) in normal but not in CF airway epithelium^{4,5}. So when the mutant gene was found to encode an ABC transporter homologue, with several consensus sites for phosphorylation by PKA, its product was suggested to regulate chloride channels by pumping out a cytoplasmic inhibitor: hence the name, CFTR². But there were more surprises. CFTR itself was shown to be a chloride channel, directly activated by phosphorylation by PKA (Fig. 1). This channel allowed chloride ions to flow equally readily in either direction, though at modest rates (conductance ≤ 10 pS; for example, see ref. 6). However, the regulated channels previously studied in airway epithelia^{4,5} conducted chloride ions rapidly (conductance ≥ 30 pS) and preferentially into the cell, and so were not CFTR channels. How, and even whether, the function of these and other channels and transporters is modulated by interaction with CFTR is still unsettled^{7,8}.

Although two-thirds of all CF disease cases can be attributed to a single mutation in CFTR, deletion of Phe 508 in NBD1, over 1,000 CF-disease-associated mutations of this 1,480-amino-acid protein have been logged (<http://www.genet.sickkids.on.ca/cftr/>). $\Delta F508$ -CFTR protein fails to mature properly and is tagged for degradation⁹. But cell incubation at room, rather than blood, temperature leads to maturation, trafficking and insertion of $\Delta F508$ CFTR into the surface membrane, where it functions nearly as well as wild type. Much effort is therefore being directed to identifying small molecules that might chaperone $\Delta F508$ to the cell surface in CF patients. Other CF-causing mutations lead to truncated protein or to full-length channels with reduced chloride-transport capacity because they either spend a smaller fraction of time open, or when open, conduct chloride ions more slowly than normal. Most disease-causing mutations have yet to be functionally characterized.

CF is a pleiotropic disease affecting all exocrine epithelia. Morbidity and mortality are largely due to chronic bacterial infection and inflammation in the lung. Inhaled DNase can ameliorate thickening of airway

¹Laboratory of Cardiac/Membrane Physiology, The Rockefeller University, New York, NY 10021, USA; ²Department of Pharmacology, University College London, London WC1E 6BT, UK; ³Department of Medical Biochemistry, Semmelweis University, 1444 Budapest, Pf. 262, Hungary.

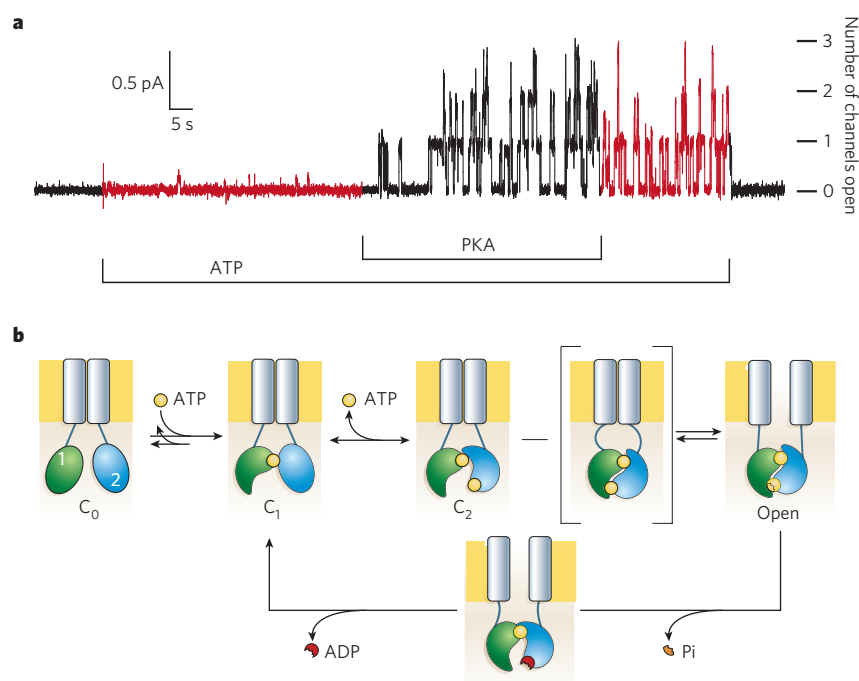


Figure 1 | Opening and closing of CFTR channels.

a, CFTR's regulatory (R) domain must be phosphorylated by PKA before ATP is able to support channel opening. The recording shows chloride current flow through individual CFTR channels upon opening. Endogenous membrane-attached phosphatases partly dephosphorylate the R domain on PKA withdrawal, reducing the probability of finding a channel open. But channels do not stop opening until ATP is removed; numbers of simultaneously open channels are indicated at the right. **b**, Present structural interpretation of ATP-dependent gating cycle of phosphorylated CFTR channels. The R domain is omitted. ATP (yellow) remains tightly bound to NBD1 (green). Walker motifs for several minutes, during which time many closed-open-closed gating cycles occur. ATP binding to NBD2 (blue) is followed by a slow channel opening step (C₂-to-Open) that proceeds through a transition state (square brackets) in which the intramolecular NBD1–NBD2 tight heterodimer is formed but the transmembrane pore (grey rectangles) has not yet opened. The relatively stable open state becomes destabilized by hydrolysis of the ATP bound at the NBD2 composite catalytic site and loss of the hydrolysis product, inorganic phosphate (Pi). The ensuing disruption of the tight dimer interface leads to channel closure.

mucus by aggregated DNA from dead cells. Although the underlying pathophysiology of CF is still debated¹⁰, the simplest hypothesis is that impaired anion movement is the primary cause. In the handful of cases examined, severity of disease seemed related to impairment of channel function. Therapeutic efforts are therefore centred on restoring the activity of mutant CFTR or enhancing alternative pathways for anion flow. Gene therapy trials to introduce CFTR DNA into airway epithelial cells began over a decade ago, but difficulties with gene delivery and expression¹¹ have led to refocusing of efforts to rescue $\Delta F508$ CFTR.

Overview of CFTR structure and function

As ABC proteins occupy up to 5% of each prokaryotic genome sequenced, there are thousands of examples of them, although few have been studied in any depth. They conform to a modular architecture, thought to comprise two transmembrane domains (TMDs), with typically six membrane-spanning α -helices each, and two cytoplasmic NBDs. Individual bacterial genes often encode single domains or fused pairs of domains (TMD–TMD, NBD–NBD or TMD–NBD), but three-quarters of the 48 human ABC genes (assigned to seven subfamilies) encode full-length ABC proteins¹². In the case of CFTR, a unique regulatory (R) domain, the target of PKA phosphorylation, links two homologous halves, giving the overall domain organization TMD1–NBD1–R–TMD2–NBD2 (ref. 2).

A common assertion, rarely tested, is that a functional ABC protein contains two TMDs and two NBDs, as found in crystal structures of the bacterial ABC transporters MsbA¹³ and BtuCD¹⁴ and in a low-resolution structure of the multidrug transporter P-gp from two-dimensional (2D) cryo-electron crystallography¹⁵. For CFTR, this means a single polypeptide, and functional measurements have addressed this question. Two CFTR polypeptides can associate¹⁶, at least through their carboxy-terminal PDZ-domain-binding motifs¹⁷. However, co-expression of two kinds of CFTR channels with distinct functional characteristics produced no hybrid channels¹⁸, and cysteine-modification tests showed that a single introduced cysteine contributed only once to a CFTR pore¹⁹. So each CFTR channel seems to be constructed from just one CFTR polypeptide.

It seems that negatively charged ions, such as chloride, accumulate near positively charged regions at each end of the pore in CFTR's transmembrane domain^{20,21} and flow through the pore after the opening of its gate by events initiated by ATP binding to the cytoplasmic NBDs (Figs 1, 2; N-terminal NBD1, green; C-terminal NBD2, blue). Before

events at CFTR's NBDs can trigger gating of the pore, the R domain must be phosphorylated by PKA²² (Fig. 1a). Once phosphorylated, CFTR channels need ATP to open normally; they close within 1 s of ATP withdrawal (Fig. 1a) and tend to stay shut until ATP is restored. A simple scheme linking these events to the pore gate (Fig. 1b) can account for most of the experimental evidence, although many details remain to be clarified. Once ATP binds to homologous nucleotide-interacting motifs in the two NBDs, these domains are thought to approach each other closely, sandwiching two ATPs in the NBD1–NBD2 interface. Upon this intramolecular heterodimer-like interaction, a signal would be transmitted through cytoplasmic-linking domains to open the gate in the transmembrane domain (Fig. 1b, grey). That channel-opening signal would be sustained until hydrolysis of one of the ATPs leads to disruption of the NBD1–NBD2 interface and separation of the NBDs. Loss of the signal allows the channel gate to close, terminating anion flow until ATP again binds to the NBDs.

Nucleotide-binding domains

High-resolution structures are not yet available for any full-length eukaryotic ABC protein but have been obtained not only for the prokaryotic MsbA and BtuCD entire transporters but also for several NBDs (for example, see refs 23–27). The NBD structures all reveal the same basic fold, with highly conserved sequence motifs positioned to interact with bound ATP (Fig. 2a).

In an ATP-bound NBD monomer²³, the phosphate chain of the nucleotide is cradled by a canonical P-loop, or Walker²⁸ A sequence (red, Fig. 2). This lies at the N terminus of a helix adjacent to a parallel β -sheet tipped by loops (one including the Walker B sequence (orange, Fig. 2), another including a 'switch' His (light green, Fig. 2)) that contact the γ phosphate of ATP in an F₁-ATPase-like ATP-binding core. On the other side of that helix, an extended loop arising from the middle of an ABC-specific anti-parallel β -region contains an aromatic residue against which the nucleotide base stacks. Evident lack of nucleotide interaction with the conserved 'ABC signature' sequence (consensus LSGGQ; pink, Fig. 2), some 15 Å distant in the so-called α -helical domain (also containing Phe 508; Fig. 2b), was initially puzzling. The puzzle was solved by biochemical demonstration of nucleotide-induced dimerization of bacterial NBDs²⁹, and by crystal structures that showed rotationally symmetric 'head-to-tail' NBD homodimers^{25–27} (compare with Fig. 2a). These nucleotide-bound NBD dimers enclosed two ATP molecules in interfacial composite sites, each comprising ATP-interacting motifs from the 'head'

of one monomer and signature sequence residues from the 'tail' of the other (Fig. 2a, c).

Recent crystal structures^{30–32} of NBD1 of CFTR show that it, too, conforms to the established NBD architecture, although its ABC-specific β -sheet includes a long, inserted loop, not found in other NBDs, and whose shape and position are still unclear. Assuming that NBD1 and NBD2 in CFTR approach each other in the way that bacterial NBDs do in the homodimeric crystals (as illustrated in Fig. 2c, d) allows construction of speculative homology models of CFTR (Fig. 2a, b), based on the structures of prokaryotic NBD dimers and of MsbA. These structural diagrams serve to illustrate the relative positioning of the transmembrane domains, cytoplasmic linking segments and NBDs, and concur with a low-resolution map³³ from cryo-electron crystallography of 2D crystals of CFTR. Much remains uncertain, however, including the precise structure of the TMDs, where homology between MsbA and CFTR is low. The structure and disposition of the R domain, lacking in other ABC proteins, is completely unknown and so is omitted from Fig. 2b (it must somehow connect the red spacefill residues, marked Rd).

A key distinction between homodimeric interactions seen so far in crystal structures of bacterial ABC proteins and the proposed NBD1–NBD2 interaction in CFTR is the asymmetry of the two expected interfacial nucleotide-binding sites, echoing the mere 27% identity between CFTR's NBD1 and NBD2. In contrast to the two essentially identical

catalytic sites found in bacterial NBD homodimers^{24–27}, each with its full complement of conserved consensus residues (Fig. 2c), the expected NBD1–NBD2 'heterodimeric' arrangement comprises only one consensus catalytic site (Fig. 2d). We call that site the 'NBD2 composite' site, as NBD2 contributes its nucleotide-binding residues of the Walker motifs, the residues that hold ATP in ATP-bound NBD monomers. The other interfacial site, the 'NBD1 composite' site, includes non-canonical residues from NBD1 Walker B (Ser instead of Glu) and switch (Ser instead of His) motifs, as well as from the NBD2 signature sequence (LSHGH instead of LSGGQ). Not surprisingly, as we will see below, evidence suggests that only the NBD2 composite site is catalytically competent.

The CFTR pore

Where is the pore in CFTR? Few systematic mutational studies have been attempted, making the residues that contribute to CFTR's anion pore uncertain. The focus of most attention has been the sixth putative membrane-spanning helix in TMD1 (gold helix, Fig. 2b), whose cytoplasmic extension connects TMD1 to NBD1. As befits a contributor to an anion-selective pore, transmembrane helix 6 is predicted to contain at least three basic residues, Arg 334 and Lys 335 (blue-green spacefill, Fig. 2b) towards the extracellular end and Arg 347 four helical turns closer to the cytoplasm. Although initially proposed to face the pore³⁴, Arg 347 was subsequently shown³⁵ to form a salt bridge to an Asp in transmembrane helix 8. A number of approaches have led to the conclusion that the positive charges at Arg 334 and Lys 335 promote entry of anions into an outer vestibule, and that the side chain of Thr338, one helical turn in, faces the pore. These include measurements of the ease of entering the pore and of passing through it, for a variety of monovalent anions, and measurements of the influence of covalent modification of introduced cysteines (reviewed in ref. 20).

Vast spans of the TMDs have not yet received attention (for example, helices 7 to 10), and effects on anion conduction of mutations at selected positions in helices 1 to 5 and 11 are not yet fully understood^{20,21}. Although the symmetrical arrangement of membrane-spanning helices in the paired TMDs of homodimeric ABC proteins, such as MsbA and BtuCD, imbues CFTR models based on them with comparable two-fold symmetry (see Fig. 2b), a lot more work is needed before the true disposition of transmembrane helices around CFTR's anion-selective pore can be specified.

Like most other chloride channels (see the review in this issue by Miller, p. 484), CFTR selects relatively poorly among small monovalent anions. Halide and pseudohalide anions larger than chloride, which escape their water shell more easily, also enter the channel more easily than chloride²⁰. Their lower avidity for water also keeps larger anions in the pore longer, allowing them to impair chloride-ion flow by competition²⁰. These characteristics imply a fairly featureless interior of the CFTR pore where no region makes very intimate contact with permeating chloride ions²⁰. This contrasts with the close embrace of potassium ions by rings of backbone carbonyl oxygens in the selectivity filter of potassium channels. Whereas potassium channels need to discriminate between similarly sized sodium and potassium ions, CFTR channels face the simpler task of distinguishing chloride (and probably bicarbonate) from larger anions such as phosphate and sulphate. Loss of regulated conduction of bicarbonate¹⁰ by CFTR channels may contribute to both airway and pancreatic-duct disease in CF.

Phosphorylation of the R domain

Gating of the CFTR channel is strictly dependent on R-domain phosphorylation by PKA. CFTR channel activity increases at least 100-fold upon phosphorylation³⁶ (from an open probability of <0.003 for non-phosphorylated CFTR channels in saturating [ATP]). The R domain comprises the ~200 residues linking NBD1 to TMD2 and contains multiple consensus sites for phosphorylation by PKA; up to six have been found phosphorylated in cells^{37–39} and up to eight in isolated CFTR⁴⁰. Among other kinases, both Ca²⁺-dependent and independent isoforms of protein kinase C (PKC) can phosphorylate R-domain sites^{22,38,41}. Although CFTR phosphorylation by PKC itself causes only

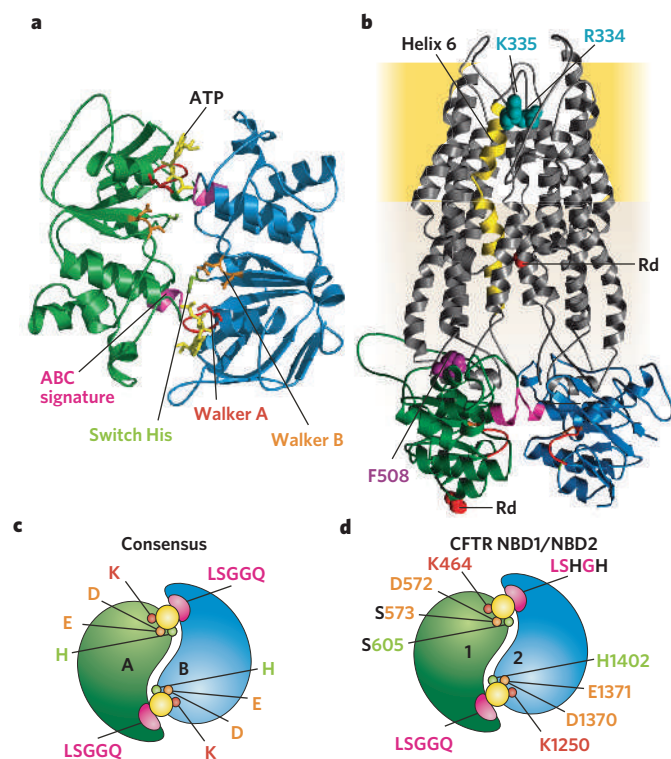


Figure 2 | Possible structure and organization of domains in CFTR.

a, Homology model of CFTR NBD heterodimer (NBD1 in green, NBD2 in blue) based on several ATP (yellow) bound NBD crystals, including human CFTR NBD1 F508A³¹ and dimeric MJ0796²⁵, MalK²⁶ and HlyB²⁷. **b**, Highly speculative homology model of CFTR including its two TMDs (grey) modelled on bacterial MsbA structures¹³. Each TMD comprises cytoplasmic segments that link membrane-embedded α -helices (one directly) to the NBDs (colour-coded as in a). Membrane-spanning helix 6 (gold) has been implicated in forming part of the anion-permeation pathway, and positively charged residues (blue-green) in it are thought to recruit anions into an outer vestibule. In CF patients, deletion of Phe 508 (purple), at the interface between NBD1 and TMD1, causes post-translational misfolding and degradation⁹. **c**, **d**, Illustrations contrasting NBD-homodimer (**c**; for example, MJ0796 and MalK) consensus catalytic site residues (for example, MJ0796 and MalK) and expected CFTR NBD1–NBD2 heterodimer (**d**) with only one consensus catalytic site.

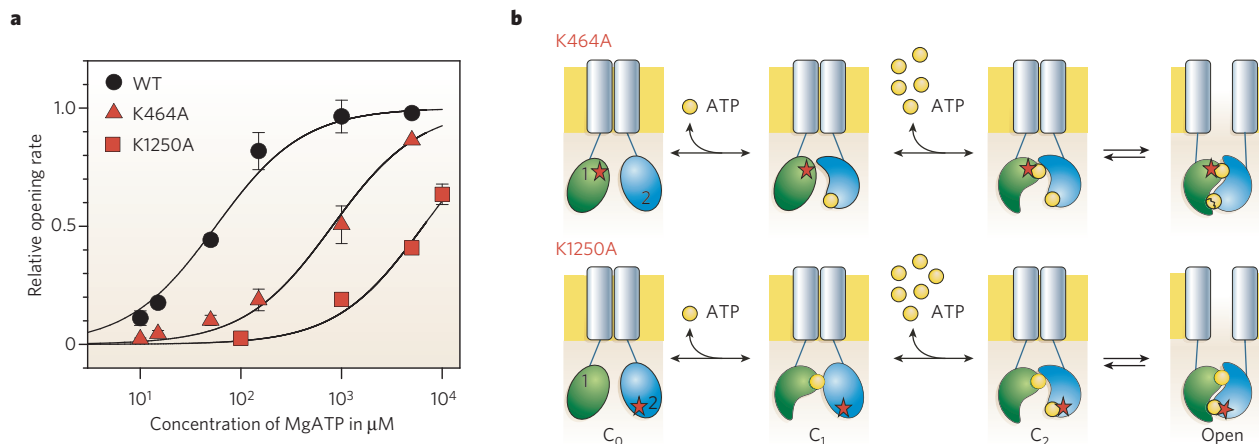


Figure 3 | The conserved Walker A lysine is critical for ATP binding in each NBD. **a**, Dependence on cytoplasmic-surface ATP concentration of rate of channel opening for wild-type CFTR (WT) and for CFTR channels with a single mutation of the NBD1 (K464) or NBD2 (K1250) Walker A lysine. Either mutation lowers the apparent affinity for ATP-dependent channel opening, implying that ATP normally binds at both sites before a channel opens. **b**, Diagram illustrating the simplest interpretation of results in **a**. At lower [ATP], channel opening is limited by low success rate of ATP binding to the composite site that harbours the mutation (red star) of the conserved lysine. ATP first binds to the non-mutant site, that is to the NBD2 site (blue) in K464A channels (upper row), but to the NBD1 site (green) in K1250A channels (lower row). When at higher concentrations, ATP also occupies the mutant site, the NBDs dimerize and the channel opens.

modest channel activation^{22,41}, it potentiates²², and might even be a pre-requisite for^{42,43}, subsequent channel activation by PKA.

The unphosphorylated R domain exerts an inhibitory influence on CFTR channel gating that can be relieved by phosphorylating it, or deleting it^{44,45}. Thus, at saturating [ATP], the open probability of CFTR channels cut in half and lacking the R domain is ~60% of that of fully phosphorylated CFTR channels similarly cut in half but retaining the entire CFTR sequence, including the R domain⁴⁵ (which all testifies to the nature of CFTR's modular ABC-protein architecture). So loss of inhibition by the unphosphorylated R domain can account for over half of the more than two orders of magnitude increase in open probability of intact wild-type CFTR channels upon phosphorylation³⁶. Any additional stimulatory effect of the phosphorylated R domain⁴⁶ is therefore probably small (less than twofold).

Just how R-domain phosphorylation controls channel gating is not clear. No other ABC protein contains an R domain, and its sequence, though conserved across species⁴⁷, shows no homology to other proteins. Its shape and location in the CFTR channel are therefore unknown. CFTR-channel activity changes gradually as increasing numbers of target serines are mutated^{37,48} or as increasing PKA concentrations incrementally phosphorylate wild-type CFTR^{39,43}. Results like these have fostered notions of redundancy among serines in a structureless R domain, and of action by accumulated negative charge^{37,49}. But this is unlikely to be the whole story. For instance, purified R-domain protein undergoes discernible conformational changes upon phosphorylation^{39,50}. In addition, two of the most readily phosphorylated serines inhibit channel activity when phosphorylated, despite undoubted introduction of negative charge^{39,51}. Although direct introduction of negative charge by substituting aspartate⁵² or glutamate⁵³ for six to eight serines does result in partly active channels, their maximal activity is 2–3-fold lower than that of phosphorylated wild-type CFTR. And circular dichroism spectra of the purified R domain indicated that the introduced acidic residues imperfectly mimic phosphoserines⁵⁴. Also, a role for a conformational change of the R domain is suggested by the finding that the prolyl-isomerase cyclophilin A stimulated wild-type CFTR channel activity, but not that of mutants lacking three conserved R-domain prolines⁵⁵.

The R domain does not seem to prevent ATP from binding at the NBDs or hydrolysis, because phosphorylation is reported not to alter photolabelling by 8-azido-ATP^{56–58}. Nor can inhibition of gating in the unphosphorylated channel be attributed to the serine-containing non-conserved structural elements in and around CFTR's NBD1^{30–32}, say by preventing NBD1–NBD2 interaction. This is because excision of neither segment vitiated the strict dependence of channel gating on

phosphorylation by PKA³⁶.

By contrast, simply severing CFTR where the R domain connects to TMD2 partly disrupts inhibition of gating in unphosphorylated channels^{45,59}. So the closed conformation of the pore might be stabilized (relative to the open conformation) by an action on the TMDs of the unphosphorylated R domain that requires its covalent attachment to transmembrane helix 7? That interaction could be relieved by R-domain phosphorylation — perhaps through its association with some other part of the channel — and the free energy difference between open- and closed-state conformations thereby lowered sufficiently to enable ATP-mediated closed-to-open transitions. Physical association of the R domain with CFTR's N terminus might be required for channel activation by PKA⁶⁰ and might be promoted by R-domain phosphorylation⁶¹.

CFTR channel gating by ATP

The location of the gate that prevents ion flow through closed CFTR channels is not known. But the machinery attached to the strings that pull the gate open and closed has recently come into view. Opening and closing of the CFTR pore are now believed to be linked to formation and disruption of a tight NBD1–NBD2 complex, driven by ATP binding and hydrolysis events (Fig. 1b). Phosphorylated CFTR channels can be opened by a broad range of nucleoside triphosphates, including ATP, GTP, ITP, UTP, CTP, AMP-CPP⁶² and 8-azido-ATP⁵⁸. But they are

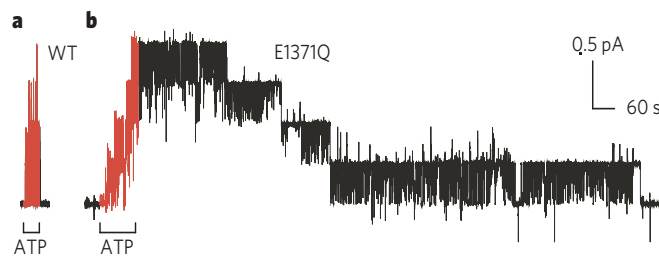


Figure 4 | ATP hydrolysis prompts channel closure. Comparison of speed of closing of wild-type CFTR (**a**) and of mutant CFTR bearing the single mutation E1371Q (**b**), of the conserved Walker B glutamate. This glutamate is required for ATP hydrolysis, but not for binding of ATP^{29,69,70,81}. The greatly delayed closing of all four E1371Q channels (**b**), compared with the four WT CFTR channels (**a**), after removal of ATP supports the conclusion that ATP hydrolysis at the NBD2 composite catalytic site controls normal channel closing^{63–66,72–77,80}, and that the channel open-burst state corresponds to NBD dimerization^{22,29,69,70,80}.

opened only inefficiently by millimolar concentrations of poorly hydrolysable analogues such as AMP-PNP, AMP-PCP or ATP- γ S^{53,62–64}, and not at all by ADP⁶². The frequency of CFTR-channel opening increases with ATP concentration along a Michaelis curve and is half maximal at about 50 μ M ATP (Fig. 3a). At saturating concentrations of ATP, a channel waits in the closed state for a second or so (depending on the temperature) before opening. This means that binding of ATP limits channel opening at low concentrations, but at saturating ATP concentrations some other slow step controls how quickly the channel can open once ATP has bound.

ATP binding and hydrolysis

Where does this rate-limiting binding of ATP occur? Mutagenesis experiments (Fig. 3) to alter the Walker motifs that contact the bound ATP in NBD crystals (Fig. 2a) confirm that, as expected, the ATP acts at CFTR's NBDs. The fact that neutralization of the Walker A Lys at either the NBD1 or the NBD2 composite site severely impairs channel opening at low ATP concentrations (Fig. 3, but compare it with ref. 65) implies that ATP binding at either site can be made rate limiting. This suggests that ATP normally occupies both sites before a channel opens^{64,66} (but compare with ref. 67).

Photolabelling of CFTR with 8-azido nucleotide, followed by separation of NBD1- and NBD2-containing fragments, confirmed binding to both NBD1 and NBD2. But only 8-azido nucleotide at NBD1 could survive washes at 0 °C before crosslinking the nucleotide with UV irradiation^{57,58,68}. NBD1 was photolabelled to a similar extent with 8-azido- $[\alpha\text{-}^{32}\text{P}]\text{-ATP}$ or 8-azido- $[\gamma\text{-}^{32}\text{P}]\text{ATP}$ ^{57,58}, even after prolonged washes at 30 °C before UV irradiation⁵⁸, whereas, even without washing, NBD2 could never be photolabelled with 8-azido- $[\gamma\text{-}^{32}\text{P}]\text{ATP}$ ⁵⁷. This all suggests that hydrolysis of 8-azido-ATP is rapid at the NBD2 composite catalytic site⁵⁷ but negligibly slow at the NBD1 composite site, even at 30 °C⁵⁸. As loss of 8-azido-ATP from NBD1 at 30 °C before photolabelling takes minutes, but opening and closing of CFTR channels in ATP takes seconds (Fig. 1a; also in 8-azido-ATP⁵⁸), nucleotide-dependent gating would seem to be controlled predominantly by events at the NBD2 composite catalytic site⁵⁸.

But how are nucleotide comings and goings at the NBD2 composite site linked to channel opening and closing? The principal consequence of mutations at the key catalytic NBD2 Walker A Lys and Walker B Asp and Glu residues (K1250, D1370 and E1371 in Fig. 2d), at saturating concentrations of ATP, is a marked increase in the time for which channels remain open (Fig. 4b). This is most easily seen from the markedly delayed step-like closure of individual mutant channels after sudden withdrawal of ATP (Fig. 4b) compared with the rapid closure of wild-type CFTR channels (Figs 1a, 4a). As biochemical measurements confirm that homologous mutations abrogate ATP hydrolysis in other ABC proteins (reviewed in refs 69, 70), and mutation K1250A has been shown to abolish ATPase activity of CFTR⁷¹, the implication is that hydrolysis of the ATP bound at the NBD2 composite catalytic site occurs before normal rapid channel closure^{64,72,73}. The idea that ATP hydrolysis sets the timing of normal CFTR-channel closure is further supported by the strong dependence of channel closing rate on Mg^{2+} concentration⁷⁴ and on temperature (activation enthalpy, $\sim 70 \text{ kJ mol}^{-1}$)^{45,75}. It is also supported by the finding that closure is markedly delayed in the presence of ATP plus the poorly hydrolysable analogue AMP-PNP^{63,76} or of ATP plus the inorganic phosphate analogue orthovanadate^{58,76,77}, which forms a stable complex with the hydrolysis product ADP after dissociation of the γ -phosphate (for example, see ref. 78).

NBD dimerization

So if hydrolysis of the NBD2-site ATP controls the timing of channel closing, what normally triggers its opening? The slow step that limits the speed of channel opening at saturating nucleotide concentration is sensitive to the structure around the γ -phosphate and, for AMP-PNP, AMP-PCP or ATP- γ S, that step occurs at $\sim 4\%$ of the maximal rate seen with ATP⁶⁴. The step is also dependent on Mg^{2+} and, at saturating ATP concentrations, in the absence of Mg^{2+} ions, occurs at only

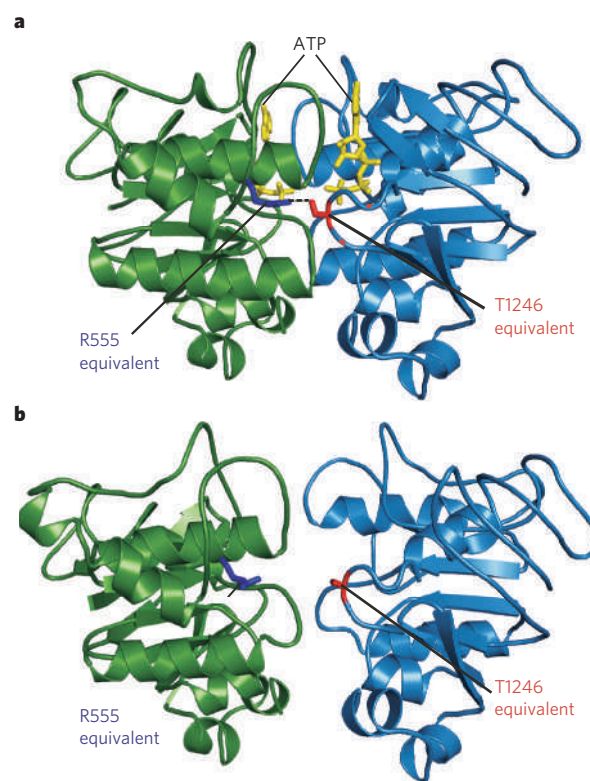


Figure 5 | Structures of the MalK homodimer. The MalK homodimer was crystallized with (a) and without (b) ATP (yellow). The structures illustrate the pincer-like motion of MalK that closes the catalytic interface upon binding of two ATP molecules²⁶. The residues corresponding to CFTR R555 (blue) and T1246 (red) form a hydrogen bond (a, dotted line) linking Walker A of one subunit with the signature sequence of the other, only in the ATP-bound crystal (a).

$<2\%$ of the maximal rate seen in their presence⁷⁴. The slow opening step is also highly temperature sensitive (activation enthalpy $\geq 100 \text{ kJ mol}^{-1}$)^{75,79}. These characteristics suggested that the process that usually opens CFTR channels is likely to be formation of an ATP-driven tight NBD1–NBD2 heterodimeric interaction⁸⁰ (Figs 1b and 2d), which is analogous to the homodimeric interaction (Figs 2a, c, 5a) found in biochemical and structural studies of bacterial NBDs^{21–24,29,69,70,81}.

One clue that the CFTR-channel open state corresponds to the heterodimerized NBD1–NBD2 conformation comes from the $\sim 1,000$ -fold stabilization of the open state (Fig. 4b) caused by the mutation E1371Q of the Walker B Glu (the possible catalytic base^{29,81}). The homologous Glu-to-Gln mutation in bacterial NBDs was required to produce the highly stable ATP-bound homodimers needed for crystallization^{25,29}. Even with that hydrolysis-impairing mutation, stable dimers were not obtained with ADP²⁹, which is consistent with a cycle of ATP-induced NBD dimerization and hydrolysis-induced dimer disruption²⁴. Opening of a CFTR channel can then be viewed as being triggered by a pincer-like drawing together of ATP-bound NBD1 and ATP-bound NBD2, as inferred from the relative positions of dimerized bacterial NBDs in the presence and absence of ATP^{26,69} (Fig. 5a, b). Channel closing would follow hydrolysis of the ATP at the NBD2 composite site, loss of the liberated phosphate and subsequent disruption of the heterodimeric NBD1–NBD2 interaction, as illustrated in Fig. 1b.

Direct evidence for a head-to-tail NBD1–NBD2 interaction in functioning CFTR channels comes from two sources. First, target cysteines introduced in pairs, one in NBD1 and the other in NBD2, can be crosslinked by short bifunctional thiol-specific reagents only when, for example, one cysteine is in a Walker motif and the other is in a signature sequence, and not when both are in Walker motifs or both are in signature sequences⁸². Second, measurements of the energetics of channel gating revealed that a residue near the NBD1 signature sequence and

another in the NBD2 Walker A motif make an energetically favourable interaction when the channels are open, or in the transition state (Fig. 1b) between closed and open states. But these residues do not interact when the channels are closed, either before or after (states C1 and C2, Fig. 1b) binding the second ATP, which is thought to trigger channel opening⁸⁰. The sign and magnitude of that coupling energy are consistent with formation of a hydrogen bond between these interacting residues as the channel opens. The corresponding two residues (labelled 'R555 equivalent' and 'T1246 equivalent' in Fig. 5b) in bacterial NBD homodimers are indeed linked by a hydrogen bond in the ATP-bound, tightly dimerized conformation^{25–27} (Fig. 5a) but not in the nucleotide-free, 'open dimer' conformation²⁶ (Fig. 5b). This strongly supports a cyclic, dynamic restructuring of the NBD1–NBD2 heterodimeric interface during the CFTR-channel gating cycle.

A bonus of the large complement of ABC-protein genes in prokaryotic genomes is that over 20,000 NBD sequences are available (<http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00005>). Statistical analysis⁸³ of these sequences suggests that the two residues that hydrogen bond across the ATP-bound NBD dimer interface (that is, the 'R555 equivalent' and 'T1246 equivalent' residues) were subject to evolutionary pressure as a pair, rather than individually. They seem to have co-evolved to retain the ability to form a hydrogen bond that precisely spaces their α -carbon atoms in the dimer (as in Fig. 5a). Analysis of the distributions of potential hydrogen-bond donor (R555 equivalent) and acceptor (T1246 equivalent) side chains shows that the two positions have changed in concert so that a shorter acceptor is more frequently paired with a longer donor and vice versa⁸⁰. The broad conservation of NBD-sequence motifs throughout prokaryotic and eukaryotic ABC proteins, and specific conservation of the structural underpinnings of the hydrogen bond found to undergo dynamic changes during the gating cycle of CFTR channels, implies that such NBD dimerization is crucial to the function of most ABC proteins.

Outlook

Although progress is being made, much remains to be learned about CFTR, including the structures and locations of the anion pore in the transmembrane domain, of its gate and of the R domain. It would also be good to know how graded R-domain phosphorylation controls communication between the NBDs and the gate. The transduction pathway by which conformational rearrangements at the NBDs are transmitted through the protein to the transmembrane domains is not known for any ABC protein. This question could be addressed by further examination of the extensive evolutionary record for ABC proteins. Statistical analysis of sequence alignments may reveal sets of positions at which amino-acid distributions are correlated. Thermodynamic mutant-cycle analysis in CFTR⁸⁰ could then be used to test whether statistical coupling reflects conservation of a network of energetically coupled residues that mediates transmission of allosteric signals. It may be possible to follow how energetic coupling between side-chain pairs changes, both as a function of their location in the protein and as a function of time during CFTR's channel-gating cycle.

Sequence analysis suggests that CFTR's asymmetric catalytic interface, with one active and one dead composite site, may be characteristic of all ABC-C subfamily members (CFTR, MRPs and SURs) but not of ABC proteins in general. The structural and mechanistic detail relevant to ABC-C proteins that can be gleaned from studies of symmetric, homodimeric ABC proteins is therefore limited. Determination of the structure of any comparably asymmetric ABC protein would greatly advance our understanding of how the ABC-C proteins CFTR, MRP and SUR work.

Much also remains to be learned about the physiological roles and functions of CFTR — for example, the identities of other proteins with which CFTR interacts (reviewed in refs 7, 8), whether CFTR functions as an ATP-driven transporter⁸⁴ and whether, and under what conditions, CFTR behaves like an adenylate kinase^{85,86}.

Present treatment of cystic fibrosis is essentially symptomatic, and although average life expectancy for CF patients is now just over 30

years, there is as yet no cure. Development of procedures for delivery of wild-type CFTR-encoding DNA using viral and non-viral (such as liposome or nanoparticle) agents or of functional CFTR protein (by means of liposomes) to affected cells continues apace¹¹. Because of the many hurdles to be overcome, parallel efforts to harness high-throughput methods to find small molecules that would allow maturation of $\Delta F508$ CFTR or enhance activity of other mutant CFTR channels are important. Because of the continuous turnover of epithelia, repetition of all these treatments will be required. On the other side of the coin, small molecule inhibitors⁸⁷ will be used in the near future against CFTR channels in the intestinal tract to combat excessive fluid loss, not least in travellers' diarrhoea. ■

- Rommens, J. M. *et al.* Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**, 1059–1065 (1989).
- Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989).
- Quinton, P. M. Chloride impermeability in cystic fibrosis. *Nature* **301**, 421–422 (1983).
- Schoumacher, R. A. *et al.* Phosphorylation fails to activate chloride channels from cystic fibrosis airway cells. *Nature* **330**, 752–754 (1987).
- Li, M. *et al.* Cyclic AMP-dependent protein kinase opens chloride channels in normal but not cystic fibrosis airway epithelium. *Nature* **331**, 358–360 (1988).
- Bear, C. E. *et al.* Purification and functional reconstitution of the cystic fibrosis transmembrane conductance regulator (CFTR). *Cell* **68**, 809–818 (1992).
- Altschuler, Y., Hodson, C. & Milgram, S. L. The apical compartment: trafficking pathways, regulators and scaffolding proteins. *Curr. Opin. Cell Biol.* **15**, 423–429 (2003).
- Li, C. & Naren, A. P. Macromolecular complexes of cystic fibrosis transmembrane conductance regulator and its interacting partners. *Pharmacol. Ther.* **108**, 208–223 (2005).
- Du, K., Sharma, M. & Lukacs, G. L. The $\Delta F508$ cystic fibrosis mutation impairs domain-domain interactions and arrests post-translational folding of CFTR. *Nature Struct. Mol. Biol.* **12**, 17–25 (2005).
- Wine, J. J. Acid in the airways. Focus on 'Hyperacidity of secreted fluid from submucosal glands in early cystic fibrosis'. *Am. J. Physiol. Cell Physiol.* **290**, C669–C671 (2006).
- Davies, J. C. & Alton, E. W. Airway gene therapy. *Adv. Genet.* **54**, 291–314 (2005).
- Dean, M., Rzhetsky, A. & Allikmets, R. The human ATP-binding cassette (ABC) transporter superfamily. *Genome Res.* **11**, 1156–1166 (2001).
- Reyes, C. L. & Chang, G. Structure of the ABC transporter MsbA in complex with ADP:vanadate and lipopolysaccharide. *Science* **308**, 1028–1031 (2005).
- Locher, K. P., Lee, A. T. & Rees, D. C. The *E. coli* BtuCD structure: A framework for ABC transporter architecture and mechanism. *Science* **296**, 1091–1098 (2002).
- Rosenberg, M. F., Callaghan, R., Modok, S., Higgins, C. F. & Ford, R. C. Three-dimensional structure of P-glycoprotein: the transmembrane regions adopt an asymmetric configuration in the nucleotide-bound state. *J. Biol. Chem.* **280**, 2857–2862 (2005).
- Ramjessingh, M., Kidd, J. F., Huan, L. J., Wang, Y. & Bear, C. E. Dimeric cystic fibrosis transmembrane conductance regulator exists in the plasma membrane. *Biochem. J.* **374**, 793–797 (2003).
- Raghuram, V., Mak, D. D. & Foskett, J. K. Regulation of cystic fibrosis transmembrane conductance regulator single-channel gating by bivalent PDZ-domain-mediated interaction. *Proc. Natl Acad. Sci. USA* **98**, 1300–1305 (2001).
- Chen, J. H., Chang, X. B., Aleksandrov, A. A. & Riordan, J. R. CFTR is a misnomer: Biochemical and functional evidence. *J. Membr. Biol.* **188**, 55–71 (2002).
- Zhang, Z. R. *et al.* Determination of the functional unit of the cystic fibrosis transmembrane conductance regulator chloride channel: One polypeptide forms one pore. *J. Biol. Chem.* **280**, 458–468 (2005).
- Dawson, D. C., Liu, X., Zhang, Z. & McCarty, N. A. in *Cystic Fibrosis Transmembrane Conductance Regulator* (eds Kirk, K. L. & Dawson, D. C.) 1–34 (Kluwer/Plenum, New York, 2003).
- Linsdell, P. Mechanism of chloride permeation in the cystic fibrosis transmembrane conductance regulator chloride channel. *Exp. Physiol.* **99**, 123–129 (2006).
- Tabcharani, J. A., Chang, X. B., Riordan, J. R. & Hanrahan, J. W. Phosphorylation-regulated Cl⁻ channel in CHO cells stably expressing the cystic fibrosis gene. *Nature* **352**, 628–631 (1991).
- Hung, L. W. *et al.* Crystal structure of the ATP-binding subunit of an ABC transporter. *Nature* **396**, 703–707 (1998).
- Hopfner, K. P. *et al.* Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily. *Cell* **101**, 789–800 (2000).
- Smith, P. C. *et al.* ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. *Mol. Cell* **10**, 139–149 (2002).
- Chen, J., Lu, G., Lin, J., Davidson, A. L. & Quirocho, F. A. A tweezers-like motion of the ATP-binding cassette dimer in an ABC transport cycle. *Mol. Cell* **12**, 651–661 (2003).
- Zaitseva, J., Jenewein, S., Jumpertz, T., Holland, I. B. & Schmitt, L. H662 is the linchpin of ATP hydrolysis in the nucleotide-binding domain of the ABC transporter HlyB. *EMBO J.* **24**, 1901–1910 (2005).
- Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. Distantly related sequences in the α - and β -subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945–951 (1982).
- Moody, J. E., Millen, L., Binns, D., Hunt, J. F. & Thomas, P. J. Cooperative, ATP-dependent association of the nucleotide-binding cassettes during the catalytic cycle of ATP-binding cassette transporters. *J. Biol. Chem.* **277**, 21111–21114 (2002).
- Lewis, H. A. *et al.* Structure of nucleotide-binding domain 1 of the cystic fibrosis transmembrane conductance regulator. *EMBO J.* **23**, 282–293 (2004).
- Lewis, H. A. *et al.* Impact of the $\Delta F508$ mutation in first nucleotide-binding domain of human cystic fibrosis transmembrane conductance regulator on domain folding and

- structure. *J. Biol. Chem.* **280**, 1346–1353 (2005).
32. Thibodeau, P. H., Brautigam, C. A., Machiusi, M. & Thomas, P. J. Side chain and backbone contributions of Phe508 to CFTR folding. *Nature Struct. Mol. Biol.* **12**, 10–16 (2005).
 33. Rosenberg, M. F. *et al.* Purification and crystallization of the cystic fibrosis transmembrane conductance regulator (CFTR). *J. Biol. Chem.* **279**, 39051–39057 (2004).
 34. Tabcharani, J. A. *et al.* Multi-ion pore behaviour in the CFTR chloride channel. *Nature* **366**, 79–82 (1993).
 35. Cotten, J. F. & Welsh, M. J. Cystic fibrosis-associated mutations at arginine 347 alter the pore architecture of CFTR. Evidence for disruption of a salt bridge. *J. Biol. Chem.* **274**, 5429–5435 (1999).
 36. Csanády, L., Chan, K. W., Nairn, A. C. & Gadsby, D. C. Functional roles of nonconserved structural segments in CFTR's NH₂-terminal nucleotide binding domain. *J. Gen. Physiol.* **125**, 43–55 (2005).
 37. Cheng, S. H. *et al.* Phosphorylation of the R domain by cAMP-dependent protein kinase regulates the CFTR chloride channel. *Cell* **66**, 1027–1036 (1991).
 38. Picciotto, M., Cohn, J., Bertuzzi, G., Greengard, P. & Nairn, A. C. Phosphorylation of the cystic fibrosis transmembrane conductance regulator. *J. Biol. Chem.* **267**, 12742–12752 (1992).
 39. Csanády, L. *et al.* Preferential phosphorylation of R-domain Serine 768 dampens activation of CFTR channels by PKA. *J. Gen. Physiol.* **125**, 171–186 (2005).
 40. Neville, D. C. A. *et al.* Evidence for phosphorylation of serine 753 in CFTR using a novel metal-ion affinity resin and matrix-assisted laser desorption mass spectrometry. *Protein Sci.* **6**, 2436–2445 (1997).
 41. Berger, H. A., Travis, S. M. & Welsh, M. J. Regulation of the cystic fibrosis transmembrane conductance regulator Cl[−] channel by specific protein kinases and protein phosphatases. *J. Biol. Chem.* **268**, 2037–2047 (1993).
 42. Jia, Y., Mathews, C. J. & Hanrahan, J. W. Phosphorylation by protein kinase C is required for acute activation of cystic fibrosis transmembrane conductance regulator by protein kinase A. *J. Biol. Chem.* **272**, 4978–4984 (1997).
 43. Chappe, V. *et al.* Phosphorylation of protein kinase C sites in NBD1 and the R domain control CFTR channel activation by PKA. *J. Physiol.* **548**, 39–52 (2003).
 44. Rich, D. P. *et al.* Effect of deleting the R domain on CFTR-generated chloride channels. *Science* **253**, 205–207 (1991).
 45. Csanády, L. *et al.* Severed channels probe regulation of gating of CFTR by its cytoplasmic domains. *J. Gen. Physiol.* **116**, 477–500 (2000).
 46. Winter, M. C. & Welsh, M. J. Stimulation of CFTR activity by its phosphorylated R domain. *Nature* **389**, 294–296 (1997).
 47. Gadsby, D. C. & Nairn, A. C. Control of CFTR channel gating by phosphorylation and nucleotide hydrolysis. *Physiol. Rev.* **79**, S77–S107 (1999).
 48. Chang, X.-B. *et al.* Protein kinase A (PKA) still activates CFTR chloride channel after mutagenesis of all 10 PKA consensus phosphorylation sites. *J. Biol. Chem.* **268**, 11304–11311 (1993).
 49. Ostedgaard, L. S., Baldrsson, O., Vermeer, D. W., Welsh, M. J. & Robertson, A. D. A functional R domain from cystic fibrosis transmembrane conductance regulator is predominantly unstructured in solution. *Proc. Natl Acad. Sci. USA* **97**, 5657–5662 (2000).
 50. Dulhanty, A. M. & Riordan, J. R. Phosphorylation by cAMP-dependent protein kinase causes a conformational change in the R domain of the cystic fibrosis transmembrane conductance regulator. *Biochemistry* **33**, 4072–4079 (1994).
 51. Wilkinson, D. J. *et al.* CFTR activation: additive effects of stimulatory and inhibitory phosphorylation sites in the R domain. *Am. J. Physiol.* **273**, L127–L133 (1997).
 52. Rich, D. R. *et al.* Regulation of the cystic fibrosis transmembrane conductance regulator Cl channel by negative charge in the R domain. *J. Biol. Chem.* **268**, 20259–20267 (1993).
 53. Aleksandrov, A. A., Chang, X., Aleksandrov, L. & Riordan, J. R. The non-hydrolytic pathway of cystic fibrosis transmembrane conductance regulator ion channel gating. *J. Physiol.* **528**, 259–265 (2000).
 54. Dulhanty, A. M., Chang, X.-B. & Riordan, J. R. Mutation of potential phosphorylation sites in the recombinant R domain of the cystic fibrosis transmembrane conductance regulator has significant effects on domain conformation. *Biochem. Biophys. Res. Commun.* **206**, 207–214 (1995).
 55. Xie, J., Zhao, J., Davis, P. B. & Ma, J. Conformation, independent of charge, in the R domain affects cystic fibrosis transmembrane conductance regulator channel openings. *Biophys. J.* **78**, 1293–1305 (2000).
 56. Travis, S. M., Carson, M. R., Ries, D. R. & Welsh, M. J. Interaction of nucleotides with membrane-associated cystic fibrosis transmembrane conductance regulator. *J. Biol. Chem.* **268**, 15336–15339 (1993).
 57. Aleksandrov, L., Aleksandrov, A. A., Chang, X. B. & Riordan, J. R. The first nucleotide binding domain of cystic fibrosis transmembrane conductance regulator is a site of stable nucleotide interaction, whereas the second is a site of rapid turnover. *J. Biol. Chem.* **277**, 15419–15425 (2002).
 58. Basso, C., Vergani, P., Nairn, A. C. & Gadsby, D. C. Prolonged nonhydrolytic interaction of nucleotide with CFTR's NH₂-terminal nucleotide binding domain and its role in channel gating. *J. Gen. Physiol.* **122**, 333–348 (2003).
 59. Ostedgaard, L. S., Rich, D. P., DeBerg, L. G. & Welsh, M. J. Association of domains within the cystic fibrosis transmembrane conductance regulator. *Biochemistry* **36**, 1287–1294 (1997).
 60. Naren, A. P. *et al.* CFTR chloride channel regulation by an interdomain interaction. *Science* **286**, 544–548 (1999).
 61. Chappe, V., Irvine, T., Liao, J., Evagelidis, A. & Hanrahan, J. W. Phosphorylation of CFTR by PKA promotes binding of the regulatory domain. *EMBO J.* **24**, 2730–2740 (2005).
 62. Anderson, M. P. *et al.* Nucleoside triphosphates are required to open the CFTR chloride channel. *Cell* **67**, 775–784 (1991).
 63. Hwang, T.-C., Nagel, G. A., Nairn, A. C. & Gadsby, D. C. Regulation of the gating of CFTR Cl channels by phosphorylation and ATP hydrolysis. *Proc. Natl Acad. Sci. USA* **91**, 4698–4702 (1994).
 64. Vergani, P., Nairn, A. C. & Gadsby, D. C. On the mechanism of MgATP-dependent gating of CFTR Cl[−] channels. *J. Gen. Physiol.* **121**, 17–36 (2003).
 65. Powe, A. C. J., Al-Nakkash, L., Li, M. & Hwang, T. C. Mutation of Walker-A lysine 464 in cystic fibrosis transmembrane conductance regulator reveals functional interaction between its nucleotide-binding domains. *J. Physiol.* **539**, 333–346 (2002).
 66. Berger, A. L., Ikuma, M. & Welsh, M. J. Normal gating of CFTR requires ATP binding to both nucleotide-binding domains and hydrolysis at the second nucleotide-binding domain. *Proc. Natl Acad. Sci. USA* **102**, 455–460 (2005).
 67. Bompadre, S. G. CFTR gating II: effects of nucleotide binding on the stability of open states. *J. Gen. Physiol.* **125**, 377–394 (2005).
 68. Szabó, K., Szakács, G., Hegedus, T. & Sarkadi, B. Nucleotide occlusion in the human cystic fibrosis transmembrane conductance regulator. *J. Biol. Chem.* **274**, 12209–12212 (1999).
 69. Davidson, A. L. & Chen, J. ATP-binding cassette transporters in bacteria. *Annu. Rev. Biochem.* **73**, 241–268 (2004).
 70. Higgins, C. F. & Linton, K. J. The ATP switch model for ABC transporters. *Nature Struct. Mol. Biol.* **11**, 918–926 (2004).
 71. Ramjeesingh, M. *et al.* Walker mutations reveal loose relationship between catalytic and channel-gating activities of purified CFTR (cystic fibrosis transmembrane conductance regulator). *Biochemistry* **38**, 1463–1468 (1999).
 72. Gunderson, K. L. & Kopito, R. R. Conformational states of CFTR associated with channel gating: the role ATP binding and hydrolysis. *Cell* **82**, 231–239 (1995).
 73. Carson, M. R., Travis, S. M. & Welsh, M. J. The two nucleotide-binding domains of cystic fibrosis transmembrane conductance regulator (CFTR) have distinct functions in controlling channel activity. *J. Biol. Chem.* **270**, 1711–1717 (1995).
 74. Dousmanis, A. G., Nairn, A. C. & Gadsby, D. C. Distinct Mg²⁺-dependent steps rate limit opening and closing of a single CFTR Cl[−] channel. *J. Gen. Physiol.* **119**, 545–559 (2002).
 75. Mathews, C. J., Tabcharani, J. A. & Hanrahan, J. W. The CFTR chloride channel: nucleotide interactions and temperature-dependent gating. *J. Membr. Biol.* **163**, 55–66 (1998).
 76. Gunderson, K. L. & Kopito, R. R. Effects of pyrophosphate and nucleotide analogs suggest a role for ATP hydrolysis in cystic fibrosis transmembrane regulator channel gating. *J. Biol. Chem.* **269**, 19349–19353 (1994).
 77. Baukowitz, T., Hwang, T.-C., Nairn, A. C. & Gadsby, D. C. Coupling of CFTR Cl channel gating to an ATP hydrolysis cycle. *Neuron* **12**, 473–482 (1994).
 78. Urbatsch, I. L., Sankaran, B., Weber, J. & Senior, A. E. P-glycoprotein is stably inhibited by vanadate-induced trapping of nucleotide at a single catalytic site. *J. Biol. Chem.* **270**, 19383–19390 (1995).
 79. Aleksandrov, A. A. & Riordan, J. R. Regulation of CFTR ion channel gating by MgATP. *FEBS Lett.* **431**, 97–101 (1998).
 80. Vergani, P., Lockless, S. W., Nairn, A. C. & Gadsby, D. C. CFTR channel opening by ATP-driven tight dimerization of its nucleotide-binding domains. *Nature* **433**, 876–880 (2005).
 81. Tomblin, G., Bartholomew, L. A., Urbatsch, I. L. & Senior, A. E. Combined mutation of catalytic glutamate residues in the two nucleotide binding domains of P-glycoprotein generates a conformation that binds ATP and ADP tightly. *J. Biol. Chem.* **279**, 31212–31220 (2004).
 82. Mense, M., Nairn, A. C. & Gadsby, D. C. CFTR chloride channel activation in *Xenopus* oocytes by forskolin/IBMX promotes formation of a Rad50-like NBD1/NBD2 dimer. *Biophys. J.* **90**, 310a (2006).
 83. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).
 84. Kogan, I. *et al.* CFTR directly mediates nucleotide-regulated glutathione flux. *EMBO J.* **22**, 1981–1989 (2003).
 85. Randak, C. O. & Welsh, M. J. Adenylate kinase activity in ABC transporters. *J. Biol. Chem.* **280**, 34385–34388 (2005).
 86. Gross, C. H. *et al.* Nucleotide binding domains of cystic fibrosis transmembrane conductance regulator, an ABC-transporter, catalyze adenylate kinase activity but not ATP hydrolysis. *J. Biol. Chem.* (in the press).
 87. Muanprasat, C. *et al.* Discovery of glycine hydrazide pore-occluding CFTR inhibitors: mechanism, structure-activity analysis, and in vivo efficacy. *J. Gen. Physiol.* **124**, 125–137 (2004).

Acknowledgements This review and our research on CFTR were supported by grants from the NIH and Fogarty International Center (to D.C.G.). We dedicate this review to our late colleague Benjamin Angel, MA, MB BS (3rd December 1978 to 2nd October 2005), whose courage and grace in the face of CF continue to inspire us.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to D.C.G. (gadsby@rockefeller.edu).

ClC chloride channels viewed through a transporter lens

Christopher Miller¹

Since its discovery, the ClC family of chloride channels has presented biophysicists with unexpected behaviours and unusual surprises. The latest of these is the realization that not only does the family feature genuine chloride channels, it also includes proton-coupled chloride transporters, which move chloride ions and protons across the membrane in opposite directions. The crystal structure of such a transporter serves as a useful platform for understanding ClC channels, and features of chloride/proton exchange-transport may provide a key for comprehending voltage-dependent gating of the channels.

Let's face it: in the world of membrane physiology and biophysics, Cl⁻ channels have long suffered as poor cousins in an aristocratic family. For more than 50 years, the study of ion channels has been dominated by these membrane proteins' neuronal functions: initiating action potentials at the synapse, propagating signals along axons and dendrites, shaping electrical waveforms and triggering neurotransmitter release through Ca²⁺ influx. These activities appeal to our sense of drama for their central position in human biology, their mechanistic intricacy at the level of protein chemistry and the reliance, for their detection, on high-resolution experimental methods. Electrical signalling in neurons almost exclusively uses Na⁺, K⁺, Ca²⁺ and H⁺ — cations — as current-carriers. Channels that employ Cl⁻, aside from their role at inhibitory synapses, have been traditionally viewed as inconspicuous molecular labourers supporting housekeeping physiologies that operate in the background, often associated with epithelial secretion of fluids such as mucus, sweat, and digestive and reproductive effluvia. These are necessary jobs, to be sure, but unglamorous ones that evoke the popular perception of Cl⁻ channels as the *Gastarbeiter* of membrane transport, quietly cleaning up the mess after the party.

In the past decade, however, Cl⁻ channels of the widespread ClC family have enjoyed some unaccustomed attention for four reasons. First, the molecular identification of the founding ClC channel¹ led swiftly to the realization that this is a large molecular family, with isoforms expressed in nearly every cell of every eukaryotic organism — the human genome has nine — and in many prokaryotes as well^{2,3}. The near-ubiquity of ClCs attests to their functional importance, as is now substantiated by the growing collection of human diseases arising from disruption of the genes encoding these proteins (Table 1). Second, the electrophysiological behaviours of ClC channels, and the structures underlying them, are idiosyncratically unlike those of the cation-conducting channels, and they present analytical challenges for which our extensive knowledge of those familiar, high-profile proteins is of little relevance. Third, high-resolution crystal structures of bacterial ClCs^{4–7} now provide a firm foundation upon which to build ideas of how these proteins work. Finally, certain ClCs that had been presumed to be channels are now recognized instead as belonging to a class of membrane 'pump' proteins^{8–10}, variously known as secondary active transporters, exchangers or antiporters. These proteins stoichiometrically exchange Cl⁻ for H⁺ through a cycle of conformational changes. This type of transporter pumps Cl⁻ uphill against a concentration gradient at the expense of H⁺ moving downhill, or vice

versa. The inescapable inference — that the ClC family comprises channel and transporter subtypes, the former residing in plasma membranes and the latter in acidifying intracellular membranes^{6,9,10} — was a big surprise. It signals that membrane proteins built along similar structural lines can support ion-transport mechanisms that are vastly different in thermodynamic essence.

This review deals with structure–mechanism relations in ClC channels rather than with their biological purposes, which have been recently summarized¹¹. An irony confronts us at the outset: although most functional work on ClCs has been done on Cl⁻ channels, the structures solved are all of the Cl⁻/H⁺ exchange-transporter subtype. How far can we stretch bacterial transporter structure to explain eukaryotic channel function? It turns out that the electrophysiological quirks of ClC channels have been dropping hints of transporter-like behaviour for the past 20 years; likewise, the newly recognized Cl⁻/H⁺ transporters display channel-like features. So the ClC family straddles a line between pumps and leaks that is customarily viewed as an absolute black-and-white divide. This review will straddle the same line to show how recent results on the ClC exchange-transporter subclass help illuminate several longstanding puzzles about the workings of ClC channels.

The bacterial ClC Cl⁻/H⁺ exchange transporter

For all types of ion channel, structure–function analysis means tackling two things: ion permeation and gating. How does the channel protein usher the favoured ion through the pore while forsaking all others? How

Table 1 | Human pathologies of ClC proteins

Isoform	Tissue, membrane distribution	Associated disease	Cell-organ phenotype of null
ClC-1 ⁵⁶	Skeletal muscle, plasma membrane	Many myopathies	Hyperexcitable muscle
ClC-2 ⁵⁷	Widespread, plasma membrane	Idiopathic epilepsy	Unknown
ClC-Kb ⁵⁸	Kidney, plasma membrane	Bartter's syndrome	Hypotension, impaired salt reuptake
ClC-5 ⁵⁹	Widespread, endosomes	Dent's disease	Chronic kidney stones
ClC-7 ⁶⁰	Widespread, endosomes, lysosomes	Osteopetrosis	Dense, fragile bones, lysosome insufficiency

¹Department of Biochemistry, Howard Hughes Medical Institute, Brandeis University, Waltham, Massachusetts 02454, USA.

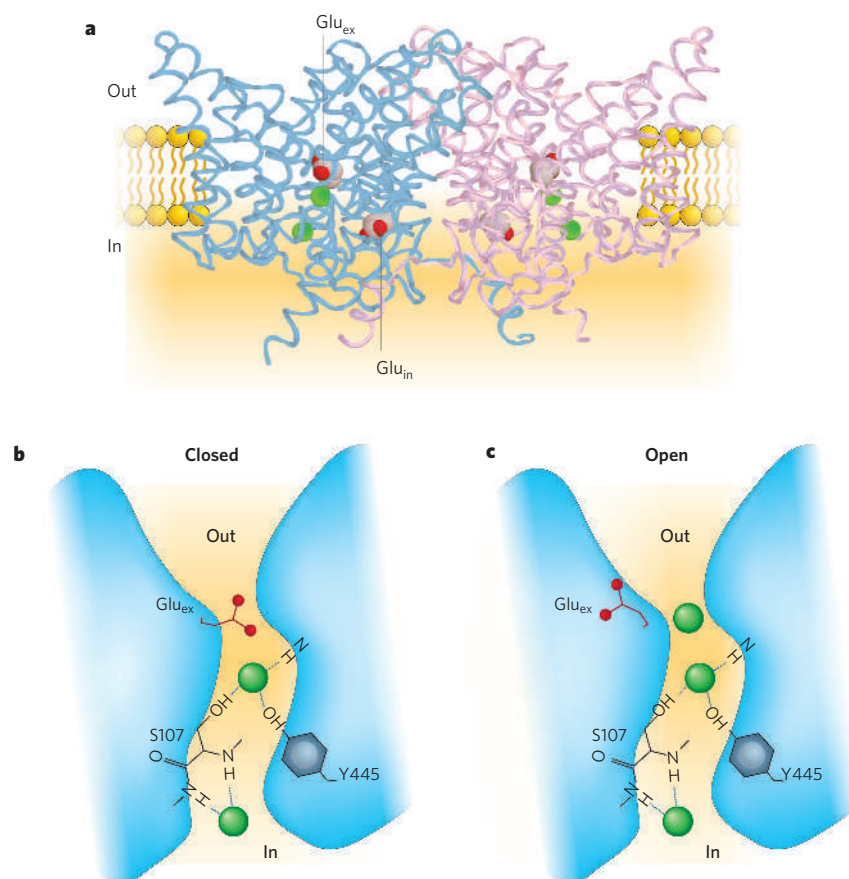


Figure 1 | Structure of ClC-ec1. **a**, The protein is depicted in a lipid membrane, with the two subunits differently coloured. Chloride ions are shown as green spheres, and the two proton-transfer glutamate residues, Glu_{ex} and Glu_{in}, are spacefilled, with carboxyl oxygen atoms in red. **b**, **c**, Diagrammatic representation of Cl[−]-binding region in 'closed' and 'open' conformations defined by the position of the side chain of Glu_{ex}. Closed conformation shows two Cl[−] ions in the 'central' and 'internal' positions; a third Cl[−] ion, in the 'external' position appears in the open conformation.

does it switch between open and closed conformations? The answers lie in the proteins' structures. Accordingly, this discussion will begin with the structure of ClC-ec1 from *Escherichia coli*^{4,5}, a transporter-type isoform. This 50-kDa protein (Fig. 1a), crystallized from detergent micelles and solved at 2.5 Å resolution, forms a homodimer with two-fold symmetry, as expected from much previous work^{12–15}. It is mostly membrane-embedded, with short cytoplasmic N and C termini. The protein is α-helical with a complicated transmembrane topology that defies hydrophobicity analysis; eighteen membrane-embedded helices are present in each subunit, eight of them in 're-entrant hairpins', where the polypeptide traverses only part of the membrane and then returns to the side from which it entered. Another notable feature is a pseudo-repeat produced by rotating the first half of the sequence 180° around an axis parallel to the membrane, so that the second half is 'upside down'. This pseudo-symmetry, also seen in aquaporins, was not predicted from, but can be rationalized after the fact by, the presence of weakly repeated sequence-elements⁴.

In striking contrast to structures of familiar ion channels, ClC-ec1 lacks a transmembrane pore. This conspicuous absence is now understood to reflect the fact that the protein is a transporter, not a channel. Buried within each subunit there is a prominent active site — an anion-coordinating region occupied by a pair of Cl[−] ions in proximity (Fig. 1b). The 'central' Cl[−]-binding site is located roughly at the membrane's mid-level, whereas the 'internal' site abuts the protein's cytoplasmic surface. The central Cl[−] is dehydrated and completely surrounded by coordinating groups from the protein, among them hydroxyls from two conserved residues, Ser 107 and Tyr 445, here denoted the 'central' Ser and Tyr. The internal Cl[−] ion, ~7 Å below the central Cl[−], is partly exposed to water and partly engulfed by protein dipolar groups. Neither lysine nor arginine directly coordinates Cl[−], but several positively charged residues located 10–15 Å away emit a strong electropositive odour in this region^{16–20}.

These structural elements are unsurprising in an anion-binding locus. But an electrostatically dissonant feature is found close to the central Cl[−]: the negatively charged carboxylate group of a glutamate residue (Glu 148,

denoted the 'external' glutamate, Glu_{ex}). This side chain sits on top of the Cl[−] ion and occludes the ion's access to the extracellular solution. Substitution with an always-neutral glutamine to mimic a protonated carboxyl results in two remarkable structural changes⁵ (Fig. 1b, c). The side chain swings outward to open a pathway to the external solution, and a third Cl[−] ion appears in precisely the location thus vacated. This is compelling evidence that the protein requires a negative charge at this site and that the Glu_{ex} carboxylate is deprotonated in the wild-type structure, which we must now view as having three, not two, anion-binding sites. This third 'external' site, hungry for an anion, can accept either a carboxylate group from deprotonated Glu_{ex} or a Cl[−] ion when Glu_{ex} is protonated. Thus, Glu_{ex} participates in three tightly linked reactions required in a Cl[−]/H⁺ exchange mechanism: Cl[−] binding, protonation and a conformational change that leads to formation of a solution-access pathway. The structure's resolution precludes absolute determination of Cl[−] occupancy, but crystallographic experiments exploiting Br[−] as a heavy-ion substitute argue that all three sites can be simultaneously occupied at physiological Cl[−] concentration⁷.

As protons are crystallographically invisible here, we have no direct information about where they go during exchange for Cl[−]. Proton transfer between the solvent and the protein interior is likely to involve dissociable residues near the aqueous interfaces²¹, mutation of which would be expected to undermine the protein's ability to handle protons. Recent work on the bacterial ClC has uncovered two such proton-transfer residues — one for each side of the membrane. The external glutamate described above is one of them; if Glu_{ex} is mutated to a neutral residue, movement of H⁺, but not of Cl[−], is completely abolished⁸. A second glutamate residue (Glu 203, denoted Glu_{in}), located near the intracellular surface and appearing only in the transporter subclass, is similarly required for H⁺ transport and appears to act as an internal H⁺-transfer site⁶. Glu_{ex} resides in the Cl[−]-binding region, but Glu_{in} is located far away from it. For this reason, it seems that towards the external side of the protein, Cl[−] and H⁺ share transport pathways, which then bifurcate somewhere near Glu_{ex} and diverge towards the internal side (Fig. 1a).

What does this structure tell us about CICs of the channel subclass? Because CIC-ec1 is not itself a channel, we must tread delicately in mapping channel properties onto the structure. I will examine correspondences between channel function and transporter structure, highlighting points of harmony and of discord and endeavouring to extract from the bacterial CIC transporter new ideas about the workings of CIC channels.

CIC Cl⁻ channels

The gross architecture of CIC channels — homodimers with a gated pore in each subunit — was inferred long ago from the single-molecule properties of the thoroughly studied muscle-type isoform CIC-0 from electric fish¹². Several other CICs display similar single-channel behaviour^{22,23}, and this 'double-barrelled' construction can be considered established for the entire family, both channels and transporters. All eukaryotic CICs (and some prokaryotic homologues) carry a large cytoplasmic C-terminal sequence that is absent in the *E. coli* protein. Within this sequence reside two copies of a 'CBS domain', a fold present in many proteins, always in pairs. Mutations here severely disrupt channel behaviour^{24,25}, and ATP binding to this part of a muscle-type CIC channel modulates its gating²⁶. A recent crystal structure of the C-terminal domain of CIC-0 (ref. 27) does not reveal its function but does suggest how this domain is arranged relative to the membrane-embedded part of the protein.

Ion permeation and selectivity

The pertinent facts regarding ion permeation in CIC channels can be quickly summarized. These channels are specific to small monovalent anions²⁸ (Cl⁻, Br⁻, I⁻, NO₃⁻, SCN⁻ but not F⁻). Cations are completely excluded, as are larger anions such as H₂PO₄⁻ and SO₄²⁻. Among permeant anions only Cl⁻ is biologically abundant (except in plants, which also handle NO₃⁻; ref. 29), so strong interanionic selectivity is not needed. Br⁻ permeates at roughly half the rate of Cl⁻, whereas I⁻ and SCN⁻, by virtue of higher avidity for the pore, are 'permeant blockers'^{30–33}. Experiments using mixed-ion conditions suggest that several ions simultaneously occupy the pore during conduction^{32–34}. Most impermeant ions do not block Cl⁻ currents even when applied at high concentrations^{33,35,36}, an 'inertness' unprecedented in any other type of ion channel.

As a pump, the bacterial CIC cannot tolerate a membrane-spanning pore. But its Cl⁻-binding region (Fig. 1b, c) presents itself as a first approximation to a Cl⁻-selective diffusion pathway, providing testable clues about ionic permeation in the channel subfamily. The region's strong electropositivity effortlessly explains the exclusion of cations as well as the anion-hunger that leads to multiple Cl⁻ occupancy and its presumed functional correlate, multi-ion conduction in the channels. Mutual repulsion experienced by Cl⁻ ions close together in the same pore may account for rapid permeation^{5,7}, although the conductances

of most CIC channels are so low that it is not clear to me that any such explanation is called for.

Strong through-space electropositivity as the origin of anion selection raises the question of why impermeant anions do not block Cl⁻ currents, especially intracellular divalent anions such as HPO₄²⁻ and SO₄²⁻, which should displace monovalents and plug up the pore through an ionic-cleansing mechanism³⁷. The CIC structure provides a plausible explanation: it lacks wide, positively charged vestibules that would offer binding sites to these larger ions. If CIC channels remain as narrow as the Cl⁻-binding region shown in the structure all along their lengths, then impermeant anions would be excluded simply by size, and the pore's electropositive character would not lead to unwanted block. This ion-selectivity strategy differs from that of K⁺, Na⁺ and Ca²⁺ channels, which have localized 'selectivity filters' — short as well as narrow — in series with wide openings to bulk solvent, and which do not use ionic size to discriminate among substrates of like charge.

The structure of the bacterial CIC, along with the strong conservation of anion-binding residues, suggests that up-close chemistry also operates on permeating Cl⁻ ions. Considering the pathological conditions required for crystallization — an outrage to physiological purists — agreement is remarkably good between Cl⁻ equilibrium dissociation constants measured crystallographically, 5–30 mM (ref. 7), and half-saturation concentrations for single-channel conductance of ~50 mM (refs 30, 33, 36). In pre-structural times, many groups mutated eukaryotic CIC channels to alter open-channel currents³⁸, and these manoeuvres, now extended in light of the structure and of homology models threaded upon it, make sense, as is thoroughly summarized by Engh and Madsen³⁹. The basic result from this mutagenesis work is that adding negative (or positive) charge to the putative pore region lowers (or raises) Cl⁻ conductance, as expected, if electropositivity here acts as an attractant for Cl⁻.

Taken together, these results imply an overall similarity in structure between the Cl⁻-binding region of the bacterial transporter and the conduction pores of CIC channels. But anion-permeation behaviour also reveals clashes with structural expectations. One of the most unsettling is that mutation of the strictly conserved central Tyr (Fig. 1b, c), which in the transporter structure directly coordinates Cl⁻ through its hydroxyl dipole, has only small effects on channel conductance^{36,40,41}. This is particularly perplexing because conservative mutation of the central Ser, which coordinates the very same Cl⁻ ion, causes a drastic lowering of conductance¹⁴. Another discordant observation is that certain aromatic organic anions, which are much too large to fit in a pore resembling the structure's Cl⁻-binding region, seem to permeate a muscle-type CIC channel⁴². Thus, the location and electrostatic nature of the structure's Cl⁻-binding region roughly mirror the CIC diffusion pathway, but the structural differences that must exist between the channels and transporters are still beyond our horizon.

CIC channel gating

CIC channels open and close by two separate processes called 'fast gating' and 'common gating'. Both are apparent by inspection of single-channel recordings (Fig. 2). The former, acting on the millisecond timescale, is a simple closed–open conformational change that occurs within each subunit independently of the other^{13,14,23,31,43,44}. The latter occludes or exposes both pores simultaneously and therefore requires communication between subunits. Indeed, mutations that alter common gating are often found near the dimer interface⁴⁵. This subject has been recently reviewed³, and I will examine only a few aspects of fast gating here.

In the most deeply studied channels, fast gating is controlled by three physiological variables: pH, Cl⁻ concentration and voltage. Opening is favoured by low pH, high Cl⁻ and depolarizing voltage (that is, inside positive). Charge movement that confers voltage dependence upon opening is somehow linked to Cl⁻ ions entering the pore from the extracellular solution^{34,35,46}, and activation by intracellular, but not extracellular, H⁺ is voltage dependent^{43,46,47}.

Where is the fast gate? The short answer: at the external glutamate. In the bacterial CIC, we have seen that the Glu_{ex} carboxylate blocks extracellular access of the Cl⁻-binding region and that upon protonation the

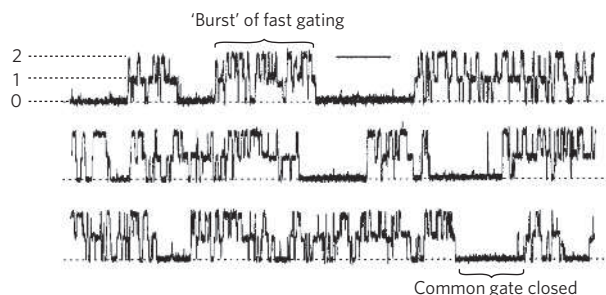


Figure 2 | A continuous single-channel recording of CIC-0. The horizontal bar over the upper trace marks 1 second. Fast gating is seen as the rapid fluctuations among three levels (representing 0, 1 or 2 pores open) within a 'burst' of activity. Common gating is the appearance and disappearance of the bursts. Holding voltage was –100 mV, and single-channel current is 0.9 pA.

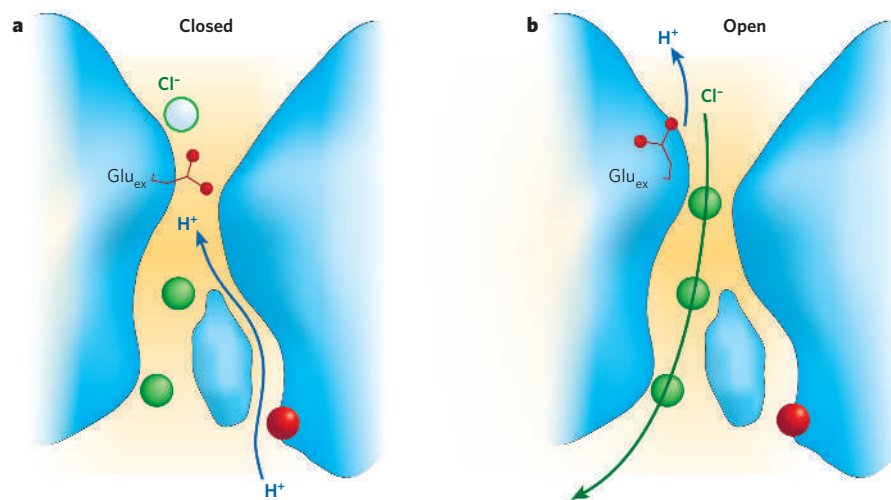


Figure 3 | A degraded-transporter model. Diagram of one subunit of a ClC channel showing a model of ClC channel fast gating. **a**, Closed state, with Glu_{ex} side chain deprotonated and blocking the pore. Allosteric Cl[−] ion occupies hypothetical site external to Glu_{ex} and thereby allows intracellular proton transfer. **b**, Open state, with protonated Glu_{ex} side chain in externally exposed position. The red circles mark the position occupied by Glu_{in} in the transporters, but which is always valine in the channels.

side chain rotates outward to open this pathway. Dutzler and colleagues⁵ tested the relevance of this small conformational change to fast gating by examining single ClC-0 channels mutated at Glu_{ex}. They found that neutralizing this residue produces almost full opening, as does exposing wild-type channels to low external pH^{5,46}. This result leads to a simple picture in which the channel, like the transporter, is occluded at its extracellular end by the negatively charged Glu_{ex} side chain, which upon protonation swings outward to open the pore.

This model explains the pH dependence of gating and, adventitiously, the constitutive activity of a ClC channel isoform in which a neutral residue replaces Glu_{ex}⁴⁸. But it does not account for voltage or Cl[−] dependence and accordingly has met with some scepticism, especially because the transporter structure shows 'gates' on both sides of the anion-binding region, with the central Ser and Tyr blocking off the cytoplasmic end. So the question naturally arises: does fast gating involve protein movements other than at Glu_{ex}? Several groups have tackled this question by focusing on the intracellular side of ClC-0. Lin and Chen⁴⁹ constructed cysteine mutants at putative pore-lining residues in order to search for positions inhibited by intracellular thiol reagents only when the channel is open. But inhibition rates remained roughly constant (within a factor of ~3) over a large voltage range, independent of whether the channel was closed or open during exposure to reagent. This result undermines the idea of a voltage-dependent gate at the pore's cytoplasmic entryway, and it accords with the failure of a designed crosslink of two helices in this area to modify fast gating³⁹. In contrast, Accardi and Pusch⁴⁰ argue that a conformational change in the intracellular region accompanies gating. Exploiting an intracellular pore-blocker that binds differentially to the closed and open channels, they showed that certain mutations near the pore's cytoplasmic end enhance the blocker's conformational preference. In my view, the conformational change at Glu_{ex} makes this residue an attractive focus for studies of the fast-gating mechanism, but as for the uniqueness of this protein movement, the outcome is still too early to call; transporters require coupled movements throughout the protein to bring about Cl[−]/H⁺ exchange, and analogous movements might also operate in the channels.

A pump-inspired proposal for ClC channel gating

For the bacterial transporter, very little is known about the Cl[−]/H⁺ exchange mechanism beyond the basic properties outlined above, including the proton-transfer roles of Glu_{in} and Glu_{ex}. But even this sparse knowledge helps to illuminate ClC channel gating. To recapitulate: in the channels, the fast gate is somehow activated by H⁺ and Cl[−], and voltage dependence seems to involve movements of these ligands from opposite sides of the membrane. This electrophysiological phenomenology provides no palpable insight into these movements, but one firm conclusion emerges from the role of Cl[−] as both permeant ion and gating activator: that gating is coupled to permeation^{34,35}. At posi-

tive voltage, for example, an extracellular Cl[−] ion that binds to the pore-associated activation site will be drawn through the pore and dissociate into the opposite solution as soon as the channel opens. In other words, gate opening is linked to the transfer of a Cl[−] ion thermodynamically downhill, a non-equilibrium process that is experimentally observable as time-asymmetric single-channel fluctuations violating microscopic reversibility⁵⁰. These phenomena have been viewed as arcane curiosities, but the ClC transporter structure endows them with fresh significance. These bizarre features of ClC channel gating (oppositely directed Cl[−] and H⁺ movements to activation sites, downhill movement of Cl[−] ions linked to opening and non-equilibrium gating) are all elements of Cl[−]/H⁺ exchange mechanisms. ClC channels plainly bear the lineaments of transporters, and for this reason, channel gating is worth revisiting with the transporter in mind.

With so little understood about Cl[−]/H⁺ exchange, drawing analogies between ClC channels and exchangers is unavoidably speculative. But I am struck by how naturally correspondences emerge. We have seen that the pores of the channel subtypes echo the transporter's anion-binding region and that Glu_{ex} in the channels acts as an extracellular gate that opens upon protonation, just as it does in the transporter. Where do the protons come from? A conventional channel-based view would hold that they must originate in the extracellular solution adjacent to Glu_{ex}. But if a transporter-ghost lurks within the channels, protons might also reach Glu_{ex} by means of 'transport' from the intracellular medium. Suppose that the channels retain in some form the transporter's proton pathway, running from the intracellular surface to meet the Cl[−]-selective pore near Glu_{ex}, and suppose further that this pathway is 'degraded', so that protons enter and move across it requiring neither mediation by Glu_{in} nor concerted Cl[−] exchange (Fig. 3). Protonation of Glu_{ex} via this route would then transfer electrical charge across the membrane and would thus confer voltage dependence upon opening. The anticipated ~tenfold enhancement of opening per 60 mV agrees with experimental values^{43,51}, and computations suggest the electrostatic plausibility of such a pathway¹⁶. In addition, this mechanism can explain, without too much shoehorning, the bewildering alterations in voltage dependence known to arise from mutations dispersed throughout the protein sequence, mutations that might disrupt the proton pathway.

This 'degraded transporter' mechanism posits Glu_{ex} as the fast gate — the sole physical blockage of the pore, as originally proposed⁵ — and it makes two rather surprising predictions. First, at high negative voltage, the channel should not close fully, but instead should asymptotically approach a non-zero open probability, the value of which rises with decreasing external pH. Glu_{ex} can be protonated from both extracellular and intracellular solutions, but only the latter reaction is voltage dependent^{43,46}. As voltage becomes very negative, the intracellular route is suppressed, but the external solution persists as a steady source of activating protons. These effects have all been observed as unexplained

experimental oddities^{35,46,51,52} and can be straightforwardly reproduced by kinetic schemes representing the degraded-transporter mechanism. A second and somewhat alarming prediction is that CIC channels cannot be ideally Cl⁻ selective. They must also conduct H⁺, because activator protons from the internal solution will dissociate into the external solvent as soon as the Glu_{ex} gate swings open — another testable, but not yet tested, example of permeation-linked gating. Proton-transit rates along this pathway need not be as fast as Cl⁻ conduction, because channels open in milliseconds whereas Cl⁻ permeates in microseconds. So protons are not required to contribute significantly to net current, but it is intriguing that an unexpected, unexplained proton flux was recently shown to accompany Cl⁻ current through a muscle-type CIC channel⁹.

A satisfactory picture of fast gating should also include activation by Cl⁻. Qualitatively, Cl⁻ activation makes sense because the permeant anion and the Glu_{ex} carboxylate compete for the same binding site in the pore and because Cl⁻ ions in the pore repel Glu_{ex} and favour gate opening. The voltage dependence of gating is currently thought to arise from movement of extracellular Cl⁻ ions into the depths of the channel protein^{34,35,46}, but this new proposal, which instead identifies H⁺ as the gating charge, assigns to Cl⁻ a different function: allosteric control of H⁺ access to Glu_{ex}. Binding of extracellular Cl⁻ to the closed channel evokes an unspecified voltage-dependent step that precedes opening^{35,46}; indeed, complete removal of extracellular Cl⁻ renders channel opening nearly independent of voltage. Let us suppose that this unknown step is simply the movement of intracellular protons to the gate; the role of Cl⁻ would then be to somehow permit this transmembrane proton transfer. When Cl⁻ concentration is low and binding sites poorly populated, proton movement through the intracellular pathway is impeded; when Cl⁻ is high, it is favoured. Linkage of Cl⁻, H⁺ and voltage in this way accounts qualitatively for the basic phenomena underlying fast gating.

But this picture has two big holes. First is the location of the Cl⁻ allosteric site, which must be on the extracellular side of the Glu_{ex} gate, where no anion is found in the transporter structure. Second is the ad hoc idea that binding of Cl⁻ to this site promotes access of intracellular H⁺ to Glu_{ex}. CIC channels offer no direct evidence for this, but a mutually reinforcing transmembrane Cl⁻–H⁺ interaction is redolent of the allosteric coupling required for exchange transport. Indeed, synergism such as this is observed in the CIC Cl⁻/H⁺ transporter itself, where binding of Cl⁻ to the anion pathway favours H⁺ transfer between Glu_{in} and Glu_{ex} by an unknown mechanism (A. Accardi, S. Lobet, C. Miller and R. Dutzler, unpublished). If an analogous effect is at work in the channels, the control of fast gating by voltage, pH and Cl⁻ could be neatly unified as an evolutionary remnant of the exchange-transport cycle.

These channel transporter analogies are rough-hewn, and I do not expect them all to survive future scrutiny. Much work is needed to test the ideas quantitatively, to elaborate details and to examine channels such as CIC-2, for instance, which gates so differently from CIC-0 and CIC-1^{53–55}. Nonetheless, a degraded-transporter mechanism, with its key element the separate pathways for Cl⁻ permeation and transmembrane H⁺ access, is a useful jumping-off place for a deeper understanding of CIC channels.

Looking forward

No subtlety is required to appreciate the questions before us or the obstacles that must be overcome to answer them. At the level of systems biology (or physiology, its unfashionable synonym), we need to know why acidification of internal compartments such as lysosomes and endosomes so determinedly employs Cl⁻/H⁺ exchangers rather than channels. At this point, we haven't a clue. Does CIC-mediated proton movement work with or against primary proton pumping by the H⁺-ATPase? Is its purpose to accumulate anions in intracellular membranes and, if so, why? Biology is sending a message here, the decoding of which will help unpack the intracellular membrane pathologies that underlie several CIC diseases. At the molecular level, a pressing aim is to crystallize and solve the structure of a proper CIC channel, so as to clarify much that is now vague, indirect and conjectural, as in this review. Finally, we need to supplement structures of the Cl⁻/H⁺ exchangers with detailed analysis of their ion-transport functions, while being alert for hints provided by

the channels. This will help us to understand a novel type of proton-coupled exchanger, and more generally to grasp the deep connections in fundamental mechanism shared in this remarkable, androgynous family of Cl⁻ transport proteins. ■

- Jentsch, T. J., Steinmeyer, K. & Schwarz, G. Primary structure of *Torpedo marmorata* chloride channel isolated by expression cloning in *Xenopus* oocytes. *Nature* **348**, 510–514 (1990).
- Pusch, M. Structural insights into chloride and proton-mediated gating of CLC chloride channels. *Biochemistry* **43**, 1135–1144 (2004).
- Chen, T.-Y. Structure and function of CLC channels. *Annu. Rev. Physiol.* **67**, 809–839 (2005).
- Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T. & MacKinnon, R. X-ray structure of a CIC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature* **415**, 287–294 (2002).
- Dutzler, R., Campbell, E. B. & MacKinnon, R. Gating the selectivity filter in CIC chloride channels. *Science* **300**, 108–112 (2003).
- Accardi, A. *et al.* Separate ion pathways in a Cl⁻/H⁺ exchanger. *J. Gen. Physiol.* **126**, 563–570 (2005).
- Lobet, S. & Dutzler, R. Ion binding properties of the CLC chloride selectivity filter. *EMBO J.* (in the press).
- Accardi, A. & Miller, C. Secondary active transport mediated by a prokaryotic homologue of CIC Cl⁻ channels. *Nature* **427**, 803–807 (2004).
- Piccolo, A. & Pusch, M. Chloride/proton antiporter activity of mammalian CLC proteins CIC-4 and CIC-5. *Nature* **436**, 420–423 (2005).
- Scheel, O., Zdebek, A. A., Lourdel, S. & Jentsch, T. J. Voltage-dependent electrogenic chloride/proton exchange by endosomal CLC proteins. *Nature* **436**, 424–427 (2005).
- Jentsch, T. J., Poet, M., Fuhrmann, J. C. & Zdebek, A. A. Physiological functions of CLC Cl⁻ channels gleaned from human genetic disease and mouse models. *Annu. Rev. Physiol.* **67**, 779–807 (2005).
- Middleton, R. E., Pheasant, D. J. & Miller, C. Purification, reconstitution, and subunit composition of a voltage-gated chloride channel from *Torpedo* electroplax. *Biochemistry* **33**, 13189–13198 (1994).
- Middleton, R. E., Pheasant, D. J. & Miller, C. Homodimeric architecture of a CIC-type chloride ion channel. *Nature* **383**, 337–340 (1996).
- Ludewig, U., Pusch, M. & Jentsch, T. J. Two physically distinct pores in the dimeric CIC-0 chloride channel. *Nature* **383**, 340–343 (1996).
- Mindell, J. A., Maduke, M., Miller, C. & Grigorieff, N. Projection structure of a CIC-type chloride channel at 6.5 Å resolution. *Nature* **409**, 219–223 (2001).
- Yin, J., Kuang, Z., Mahankali, U. & Beck, T. L. Ion transit pathways and gating in CIC chloride channels. *Proteins* **57**, 414–421 (2003).
- Faraldo-Gomez, J. D. & Roux, B. Electrostatics of ion stabilization in a CIC chloride channel homologue from *Escherichia coli*. *J. Mol. Biol.* **339**, 981–1000 (2004).
- Miloshevsky, G. V. & Jordan, P. C. Anion pathway and potential energy profiles along curvilinear bacterial CIC Cl⁻ pores: electrostatic effects of charged residues. *Biophys. J.* **86**, 825–835 (2004).
- Corry, B., O'Mara, M. & Chung, S. H. Conduction mechanisms of chloride ions in CIC-type channels. *Biophys. J.* **86**, 846–860 (2004).
- Cohen, J. & Schulten, K. Mechanism of anionic conduction across CIC. *Biophys. J.* **86**, 836–845 (2004).
- Adelroth, P. & Brzezinski, P. Surface-mediated proton-transfer reactions in membrane-bound proteins. *Biochim. Biophys. Acta* **1655**, 102–115 (2004).
- Saviane, C., Conti, F. & Pusch, M. The muscle chloride channel CIC-1 has a double-barreled appearance that is differentially affected in dominant and recessive myotonia. *J. Gen. Physiol.* **113**, 457–468 (1999).
- Weinreich, F. & Jentsch, T. J. Pores formed by single subunits in mixed dimers of different CLC chloride channels. *J. Biol. Chem.* **276**, 2347–2353 (2001).
- Estevez, R., Pusch, M., Ferrer-Costa, C., Orozco, M. & Jentsch, T. J. Functional and structural conservation of CBS domains from CLC channels. *J. Physiol. (Lond.)* **557**, 363–378 (2004).
- Hebeisen, S. & Fahlke, C. Carboxy-terminal truncations modify the outer pore vestibule of muscle chloride channels. *Biophys. J.* **89**, 1710–1720 (2005).
- Bennetts, B. *et al.* Cytoplasmic ATP-sensing domains regulate gating of skeletal muscle CIC-1 chloride channels. *J. Biol. Chem.* **280**, 32452–32458 (2005).
- Meyer, S. & Dutzler, R. Crystal structure of the cytoplasmic domain of the chloride channel CIC-0. *Structure* **14**, 299–307 (2006).
- Fahlke, C. Ion permeation and selectivity in CIC-type chloride channels. *Am. J. Physiol. Renal Physiol.* **280**, F748–F757 (2001).
- Geelen, D. *et al.* Disruption of putative anion channel gene *AtClC-a* in *Arabidopsis* suggests a role in the regulation of nitrate content. *Plant J.* **21**, 259–267 (2000).
- White, M. M. & Miller, C. Probes of the conduction process of a voltage-gated chloride channel from *Torpedo* electroplax. *J. Gen. Physiol.* **78**, 1–19 (1981).
- Miller, C. Open-state substructure of single chloride channels from *Torpedo* electroplax. *Phil. Trans. R. Soc. Lond. B* **299**, 401–411 (1982).
- Fahlke, C., Durr, C. & George, A. L. Jr. Mechanism of ion permeation in skeletal muscle chloride channels. *J. Gen. Physiol.* **110**, 551–564 (1997).
- Rychkov, G. Y., Pusch, M., Roberts, M. L., Jentsch, T. J. & Bretag, A. H. Permeation and block of the skeletal muscle chloride channel, CIC-1, by foreign anions. *J. Gen. Physiol.* **111**, 653–665 (1998).
- Pusch, M., Ludewig, U., Rehfeldt, A. & Jentsch, T. J. Gating of the voltage-dependent chloride channel CIC-0 by the permeant anion. *Nature* **373**, 527–531 (1995).
- Chen, T.-Y. & Miller, C. Nonequilibrium gating and voltage dependence of the CIC-0 Cl⁻ channel. *J. Gen. Physiol.* **108**, 237–250 (1996).
- Chen, M. F. & Chen, T. Y. Side-chain charge effects and conductance determinants in the pore of CIC-0 chloride channels. *J. Gen. Physiol.* **122**, 133–145 (2003).
- Miller, C. Ionic hopping defended. *J. Gen. Physiol.* **113**, 783–787 (1999).
- Jentsch, T. J., Stein, V., Weinreich, F. & Zdebek, A. A. Molecular structure and physiological function of chloride channels. *Physiol. Rev.* **82**, 503–568 (2002).
- Engh, A. M. & Maduke, M. Cysteine accessibility in CIC-0 supports conservation of the CIC intracellular vestibule. *J. Gen. Physiol.* **125**, 601–617 (2005).

40. Accardi, A. & Pusch, M. Conformational changes in the pore of CLC-0. *J. Gen. Physiol.* **122**, 277–294 (2003).
41. Estevez, R., Schroeder, B. C., Accardi, A., Jentsch, T. J. & Pusch, M. Conservation of chloride channel structure revealed by an inhibitor binding site in CLC-1. *Neuron* **38**, 47–59 (2003).
42. Rychkov, G. Y., Pusch, M., Roberts, M. L. & Bretag, A. H. Interaction of hydrophobic anions with the rat skeletal muscle chloride channel CLC-1: effects on permeation and gating. *J. Physiol. (Lond.)* **530**, 379–393 (2001).
43. Hanke, W. & Miller, C. Single chloride channels from *Torpedo* electroplax: activation by protons. *J. Gen. Physiol.* **82**, 25–45 (1983).
44. Ludewig, U., Pusch, M. & Jentsch, T. J. Independent gating of single pores in CLC-0 chloride channels. *Biophys. J.* **73**, 789–797 (1997).
45. Duffield, M., Rychkov, G., Bretag, A. & Roberts, M. Involvement of helices at the dimer interface in CLC-1 common gating. *J. Gen. Physiol.* **121**, 149–161 (2003).
46. Chen, M. F. & Chen, T. Y. Different fast-gate regulation by external Cl^- and H^+ of the muscle-type CLC chloride channels. *J. Gen. Physiol.* **118**, 23–32 (2001).
47. Rychkov, G. Y. *et al.* Concentration and pH dependence of skeletal muscle chloride channel CLC-1. *J. Physiol. (Lond.)* **497**, 423–435 (1996).
48. Waldegger, S. & Jentsch, T. J. Functional and structural analysis of CLC-K chloride channels involved in renal disease. *J. Biol. Chem.* **275**, 24527–24533 (2000).
49. Lin, C. W. & Chen, T. Y. Probing the pore of CLC-0 by substituted cysteine accessibility method using methane thiosulfonate reagents. *J. Gen. Physiol.* **122**, 147–159 (2003).
50. Richard, E. A. & Miller, C. Steady-state coupling of ion-channel conformations to a transmembrane ion gradient. *Science* **247**, 1208–1210 (1990).
51. Accardi, A. & Pusch, M. Fast and slow gating relaxations in the muscle chloride channel CLC-1. *J. Gen. Physiol.* **116**, 433–444 (2000).
52. Ludewig, U., Jentsch, T. J. & Pusch, M. Analysis of a protein region involved in permeation and gating of the voltage-gated *Torpedo* chloride channel CLC-0. *J. Physiol. (Lond.)* **498**, 691–702 (1997).
53. Jordt, S.-V. & Jentsch, T. J. Molecular dissection of gating in the CLC-2 chloride channel. *EMBO J.* **16**, 1582–1592 (1997).
54. Zuniga, L. *et al.* The voltage-dependent CLC-2 chloride channel has a dual gating mechanism. *J. Physiol. (Lond.)* **555**, 671–682 (2004).
55. de Santiago, J. A., Nehrkke, K. & Arreola, J. Quantitative analysis of the voltage-dependent gating of mouse parotid CLC-2 chloride channel. *J. Gen. Physiol.* **126**, 591–693 (2005).
56. Steinmeyer, K. *et al.* Inactivation of muscle chloride channel by transposon insertion in myotonic mice. *Nature* **354**, 304–308 (1991).
57. Haug, K. *et al.* Mutations in CLCN2 encoding a voltage-gated chloride channel are associated with idiopathic generalized epilepsies. *Nature Genet.* **33**, 527–532 (2003).
58. Simon, D. B. *et al.* Mutations in the chloride channel gene *CLCNKB* cause Bartter's syndrome type III. *Nature Genet.* **17**, 171–178 (1997).
59. Steinmeyer, K., Schwappach, B., Bens, M., Vandewalle, A. & Jentsch, T. J. Cloning and functional expression of rat CLC-5, a chloride channel related to kidney disease. *J. Biol. Chem.* **270**, 31172–31177 (1995).
60. Kornak, U. *et al.* Loss of the CLC-7 chloride channel leads to osteopetrosis in mice and man. *Cell* **104**, 205–215 (2001).

Note added in proof The idea of voltage dependence of fast gating arising from intracellular proton movement was also recently suggested in studies of mutations at Glu₆₈ in CLC-0 (Raverso, S., Zifarelli, G., Aiello, R. & Pusch, M. *J. Gen. Physiol.* **127**, 51–65, 2006).

Acknowledgments I thank A. Accardi, T.-Y. Chen, R. Dutzler, Z. Lu and C. Thoreen for comments and criticisms on this manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to the author (cmiller@brandeis.edu).

Southern Ocean sea-ice extent, productivity and iron flux over the past eight glacial cycles

E. W. Wolff¹, H. Fischer², F. Fundel², U. Ruth², B. Twarloh², G. C. Littot¹, R. Mulvaney¹, R. Röthlisberger¹, M. de Angelis³, C. F. Boutron³, M. Hansson⁴, U. Jonsell⁴, M. A. Hutterli^{1,5}, F. Lambert⁵, P. Kaufmann⁵, B. Stauffer⁵, T. F. Stocker⁵, J. P. Steffensen⁶, M. Bigler^{5,6}, M. L. Siggaard-Andersen⁶, R. Udisti⁷, S. Becagli⁷, E. Castellano⁷, M. Severi⁷, D. Wagenbach⁸, C. Barbante^{9,10}, P. Gabrielli¹⁰ & V. Gaspari⁹

Sea ice and dust flux increased greatly in the Southern Ocean during the last glacial period. Palaeorecords provide contradictory evidence about marine productivity in this region, but beyond one glacial cycle, data were sparse. Here we present continuous chemical proxy data spanning the last eight glacial cycles (740,000 years) from the Dome C Antarctic ice core. These data constrain winter sea-ice extent in the Indian Ocean, Southern Ocean biogenic productivity and Patagonian climatic conditions. We found that maximum sea-ice extent is closely tied to Antarctic temperature on multi-millennial timescales, but less so on shorter timescales. Biological dimethylsulphide emissions south of the polar front seem to have changed little with climate, suggesting that sulphur compounds were not active in climate regulation. We observe large glacial-interglacial contrasts in iron deposition, which we infer reflects strongly changing Patagonian conditions. During glacial terminations, changes in Patagonia apparently preceded sea-ice reduction, indicating that multiple mechanisms may be responsible for different phases of CO₂ increase during glacial terminations. We observe no changes in internal climatic feedbacks that could have caused the change in amplitude of Antarctic temperature variations observed 440,000 years ago.

The late Quaternary period is characterized particularly by strong 100,000-year (100 kyr) cycles. Ice cores play a key part in providing an understanding of climate variability during this period because many components of forcing and feedback, including greenhouse gases¹, are represented in a single core. The ionic chemistry in ice cores is mainly representative of aerosol, and has been interpreted as giving information about important environmental features such as sea-ice extent, marine biological productivity of the nearby ocean, aridity of the surrounding continents, and transport strength.

For Antarctica, both dust (representing terrestrial aerosol) and sodium (representing marine aerosol) data have been presented¹ from the Vostok core, extending 420 kyr into the past. The concentrations of both components are significantly increased in glacial periods compared to interglacials. This was attributed to a range of causes, with a particular emphasis on increased meridional transport. A significant increase in methanesulphonic acid concentrations² was attributed to increased marine biological productivity in the Southern Ocean. However, more-detailed profiles from other sites, along with a better knowledge of transport, deposition and post-depositional processes^{3–5}, have called these interpretations into question.

The climate of Antarctica over the longer period of 740,000 yr (740 kyr, to marine isotope stage (MIS)18.4) has recently been

characterized through a new ice-core record⁶. The core, drilled by the European Project for Ice Coring in Antarctica (EPICA) came from Dome C (75° 06' S, 123° 21' E, altitude 3,233 m above sea level). The deuterium (temperature proxy) record (Fig. 1) highlighted the change in glacial–interglacial temperature amplitude that occurred around 440 kyr BP (before present); interglacials in the earlier part of the record were much cooler than recent interglacials, despite similar external forcing. Here we present the chemical data only for the main oceanic (sea salt and marine biogenic) and terrestrial (including iron) aerosol components reaching Antarctica. We make new interpretations of the environmental changes during several glacial–interglacial cycles that they represent. We also investigate how the relationship between different variables altered when the climate pattern changed around 440 kyr BP.

Defining the ice-core proxies

All chemical concentrations (see Methods) are higher during glacial periods than during interglacial periods (Fig. 1). However, because of the very low snow accumulation rates at Dome C, the dominant process for aerosol deposition is almost certainly dry deposition⁷. In such a situation, the flux rather than the concentration in ice is expected to be the measure that is indicative of changes in atmospheric concentration. Part of the concentration contrast seen is

¹British Antarctic Survey, High Cross, Madingley Road, Cambridge, CB3 0ET, UK. ²Alfred Wegener Institute for Polar and Marine Research, Columbusstrasse, D-27568 Bremerhaven, Germany. ³Laboratoire de Glaciologie et Géophysique de l'Environnement, BP96, 38402 St Martin d'Hères cedex, France. ⁴Department of Physical Geography and Quaternary Geology, Stockholm University, 106 91 Stockholm, Sweden. ⁵Climate and Environmental Physics, Physics Institute, University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland. ⁶Department of Geophysics, Juliane Maries Vej 30, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark. ⁷Department of Chemistry—Analytical Chemistry Section, Scientific Pole—University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Florence, Italy. ⁸Institut für Umweltphysik, University of Heidelberg, INF 229, 69120 Heidelberg, Germany. ⁹Environmental Sciences Department, University of Venice, Calle Larga S. Marta 2137, ¹⁰Institute for the Dynamics of Environmental Processes—CNR, University of Venice, I-30123 Venice, Italy.

therefore simply due to the fact that the snow accumulation rate is estimated to be more than a factor of two higher in warm interglacials than in the coldest part of each glacial. In Fig. 2, fluxes have been calculated (using the estimated snow accumulation rate⁶) and they are used for the later discussions (see also Supplementary Information).

Iron (Fe) is derived almost entirely from continental terrestrial sources. However, we have more continuous data available for calcium (Ca), which is often used to define terrestrial sources, but which also has a marine source. Sodium (Na) has a mainly marine source, but some continental dust influence. For the remaining figures, we therefore calculated sea-salt Na (ssNa) and non-sea-salt Ca (nssCa) concentrations and fluxes (see Methods). Despite differences in detail, the good general agreement between nssCa and Fe profiles (Fig. 2) suggests that each is suitable as a terrestrial marker element at Dome C, although Fe might be preferred in discussions about iron fertilization of the ocean^{8,9}.

Methanesulphonate (MS^-) and sulphate (SO_4^{2-}) in Antarctica are both considered to be derived mainly from marine biogenic emission of dimethylsulphide (DMS)¹⁰, followed by oxidation in the atmosphere. MS^- has no other significant source for Antarctica. SO_4^{2-} comes also from sea salt, anthropogenic sources (though these have a very limited influence in Antarctica¹¹), terrestrial dust (also limited), and from volcanism—the latter causing intermittent spikes and a continuous background. The sea-salt source of sulphate can be removed using Na data (see Methods) to give non-sea-salt SO_4^{2-} (nss SO_4^{2-}). Estimates of the annual downward transport from the stratosphere to the troposphere¹¹, comparisons of the stratospheric

aerosol burden in non-volcanic and volcanic periods, and modelling studies¹⁰ suggest that the volcanic input during background periods is <10% of the deposition of SO_4^{2-} to the Antarctic ice sheet. Volcanic input during eruptive events¹² can dominate the SO_4^{2-} budget for 1–3 yr, but fills only a small time fraction, representing (for example) only about 6% of the Holocene sulphate budget¹³. In deeper ice, these spikes become diffuse, and eventually cannot be identified against the background: as a result there is no objective way to remove their influence. We have therefore averaged the data without removing the spikes or background, but can assume that the nss SO_4^{2-} flux is strongly dominated by marine biogenic sources throughout the record.

Very large glacial–interglacial contrasts in MS^- (seen also at Dome C) have been widely quoted in the past as evidence² of increased glacial-period marine biogenic activity in the Southern Ocean. However, recent evidence^{3,4,14} has shown clearly that, under modern conditions at sites with very low snow accumulation rates, MS^- concentration falls rapidly in the upper metres of snow; this is also seen in snowpits at Dome C. This is probably due to a slow evaporative loss of methanesulphonic acid, in a process analogous to that observed for nitrate¹⁵. In glacial periods, much more terrestrial material is present in the atmosphere and snow. Nitrate and, if

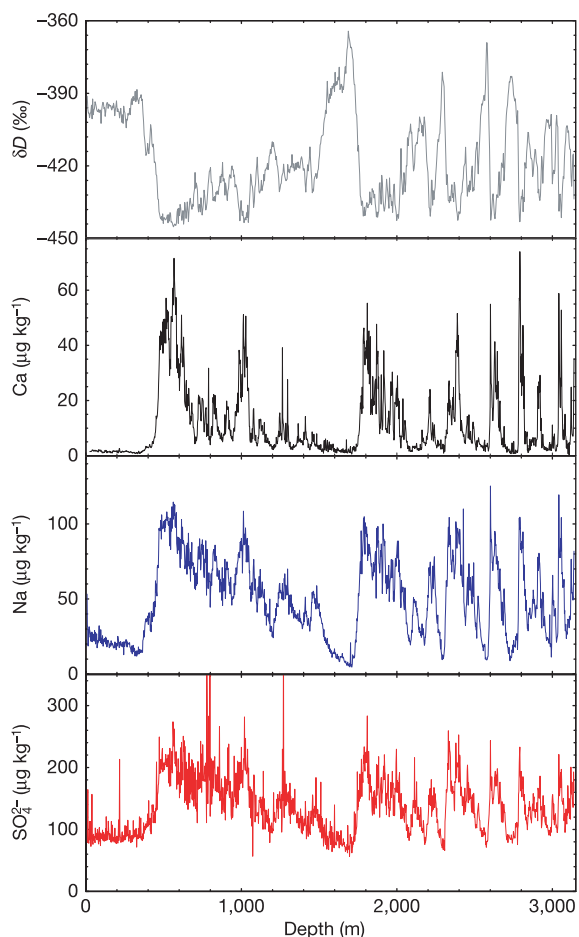


Figure 1 | Measured concentrations from the EPICA Dome C ice core. Data are on an ice depth scale. δD are averaged over 3.85-m sections⁶; chemical concentrations are averaged over 2.2-m sections.

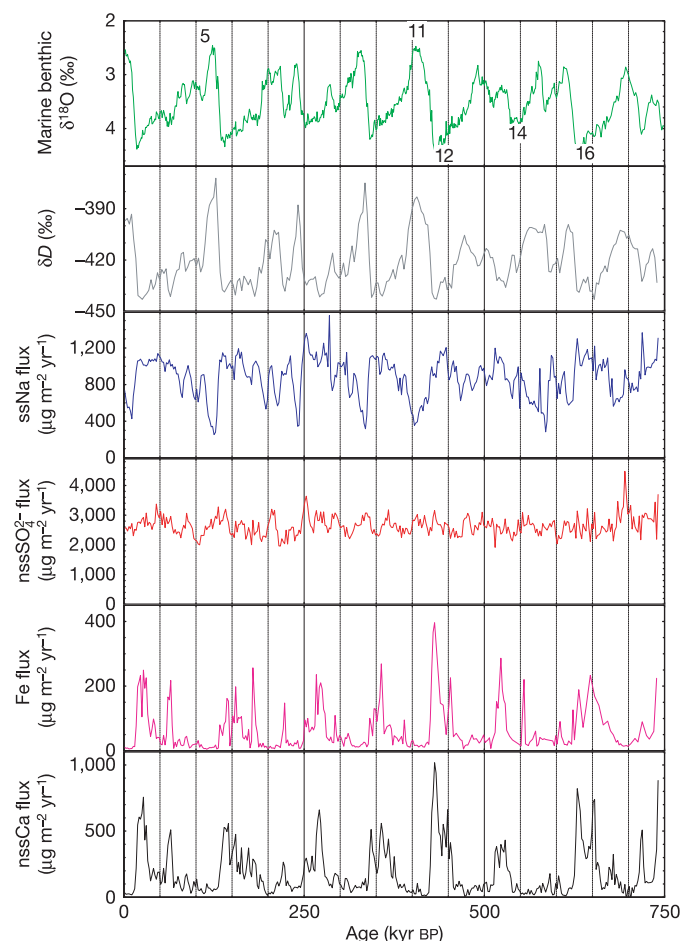


Figure 2 | Chemical measurements from the EPICA Dome C core, on an age scale. Chemical components (2-kyr averages) and δD (3-kyr averages)⁶ are from the EDC core; oxygen isotope values (1 kyr averages) from the marine benthic stack (on the LR04 timescale and with selected MIS numbers shown)³⁹. While nssCa data represent almost continuous averages within each 2-kyr period, Fe data are averages of a few discrete samples. In the deeper ice, many 2-kyr periods have no Fe data, or are based on a single data point representing only a few decades. The obvious timing mismatch between EDC and the benthic stack around MIS14 is not yet resolved³⁹.

the analogy holds, MS^- , are no longer mainly present as relatively volatile acids, but are instead preserved as non-volatile neutral salts. Therefore the higher concentrations of both ions in snow from cold periods reflects better fixation (or a different deposition mechanism), and does not indicate the past production of DMS. We therefore use nssSO_4^{2-} , which suffers no post-depositional losses, as our marker of marine biogenic emissions.

Sea ice based on sea salt

The ssNa (representing sea salt) flux has minima in interglacials, and maxima in glacial periods. The maximum values are about double the present-day value, while the minimum 2-kyr average observed is about half the present value. There is a very close relationship between Na flux and temperature throughout the 740-kyr period, with cold temperatures leading to higher Na flux (Fig. 3).

Previously, the increased cold period ssNa flux or concentration has been attributed to greater cyclonic activity at the (assumed open-water) source, and to greater meridional transport of sea salt into the continent¹. Recently, this interpretation has been challenged on several grounds¹⁶. The chemical signature of both aerosol and snow, showing a characteristic depletion of $\text{SO}_4^{2-}/\text{Na}$ compared to sea water, suggests that new sea-ice surfaces are the major source of sea salt to coastal Antarctica^{17,18}. Sea salt concentrations, even in central Antarctica, peak seasonally when sea ice is at its maximum, and there is no evidence that cyclonic activity has the seasonal contrast needed to overcome the greater travel distance this implies¹⁹. The modelled flux of sea salt arriving in Antarctica from an open-water source was found to be much lower in the last glacial maximum (LGM) compared to the present, when the data show the opposite²⁰.

The evidence for a sea-ice source of sea salt to central Antarctica¹⁶ has been circumstantial. Recently, year-round measurements of aerosol concentrations at the high-plateau inland station of Dome Fuji²¹ showed $\text{SO}_4^{2-}/\text{Na}$ ratios that were often below that expected from sea water, with a pattern exactly as found at the coastal sites. A

process of sublimation at the Antarctic snow surface was proposed²¹, but this should not lead to fractionation. The most plausible explanation is that the aerosol derives from a source depleted in SO_4^{2-} . We conclude that the sea-ice surface (including frost flowers, brine and brine-soaked snow) is probably the main source of sea salt to central Antarctica for recent winter conditions. This would imply that the sea salt flux at Dome C is related to new sea-ice production in the Indian Ocean sector of Antarctica (see Supplementary Information), at least on longer timescales. Marine diatom evidence suggests that summer sea ice was negligible in this sector even in the LGM²², which is among the coldest periods in the Dome C deuterium profile. We can conclude that annual production and maximum extent are closely related; with our interpretation, ssNa flux indicates winter sea-ice extent. In the LGM, when ssNa flux is double the present value, winter sea-ice extent is believed to have been about double that of the present²², providing an approximate calibration of our proxy.

We therefore conclude that, at multimillennial timescales, winter ice extent in this sector is very closely tied to Antarctic temperature (Fig. 3). The relationship is stable throughout the last 740 kyr, implying that the change in climatic pattern at around 440 kyr is not due to any change in the climate–sea-ice feedback. Winter ice extent has probably been considerably below the present value in recent interglacials. Sea-ice extent was intermediate between recent glacial and interglacial values during most of the weak interglacials before 440 kyr BP. The linkage between Antarctic temperature and sea-ice extent is less strong at shorter timescales; changes in sea salt and deuterium are not synchronous during major climate transitions, as discussed later. Although we recognize that our use of sea salt as a sea-ice proxy remains tentative, we suggest that the ssNa flux curve in Fig. 2 may be taken as a first attempt at a quantitative winter ice extent over the last 740 kyr for use in modelling studies. It will also be important for assessing the role of sea-ice mechanisms in CO_2 changes²³.

Marine productivity

High glacial MS^- was used to suggest increased oceanic DMS emissions during the last glacial period². This led to the suggestion that the S cycle has been very sensitive to climate change². Because the products of DMS oxidation, through their role as cloud condensation nuclei, may themselves cause climate feedbacks²⁴, the Vostok MS^- record is often quoted in discussions about climate self-regulation^{25,26}. DMS production is also quoted as one indicator of productivity, relevant to the drawdown of CO_2 , although it has to be emphasized that DMS producers may not be representative of total biological production.

In contrast to MS^- , the nssSO_4^{2-} flux is stable, within about 20%, through the entire 740-kyr record (Fig. 2). The single short-lived increase in flux, depending mainly on a single data point at about 700 kyr seems not to be replicated in a parallel analysis by the fast ion chromatography method (see Methods). The low variability in the rest of the record is remarkable, and unique among ice-core chemistry records to date. A similar result (with concentrations approximately doubled in peak glacials compared to interglacials, implying little change in flux) would have been found for Vostok² or Dome Fuji²⁷. We note that the use of concentrations rather than fluxes would have led to a different conclusion, but our assumptions about aerosol deposition leading to this choice are robust (see Supplementary Information).

The amount of nssSO_4^{2-} reaching Dome C should depend on the production and emission of DMS, on the location of emissions (and hence the transport distance), on transport speed and on transformations en route. The constant nssSO_4^{2-} flux could result from all these factors changing between glacial and interglacial (in which case their effects must fortuitously cancel out), or from the dominant factors remaining constant.

The summer sea-ice edge was close to the continent in the sector of

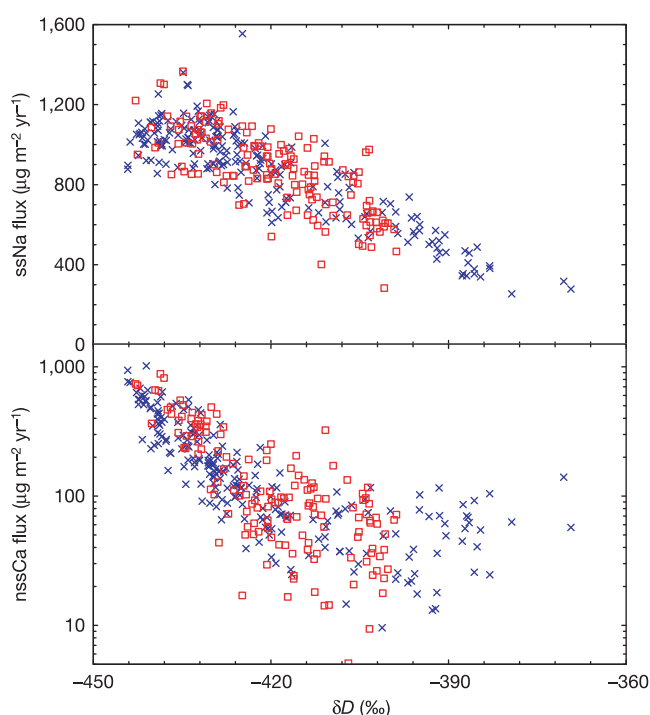


Figure 3 | Relationship between chemistry and climate. ssNa and nssCa fluxes plotted against δD for the entire ice core, using 2-kyr averages. Blue crosses are for the period from 0–440 kyr BP, red squares are for 440–740 kyr BP.

Antarctica nearest to Dome C even in the LGM²², so that the location of summer DMS emission need not have moved significantly. Model studies of tracer transport, although mainly dealing with transport from further north than the DMS source regions, suggest a rather small change in transport time for the LGM compared to the present^{28,29}. The proportion of DMS that is oxidized to SO_4^{2-} or to MS^- might vary with climate, but nssSO_4^{2-} probably remains the dominant product under all conditions³⁰. We therefore conclude that these factors are unlikely to have been decisive: allowing for some changes in these factors, and for uncertainty in our assumptions about the dominance of dry deposition and the snow accumulation rate, DMS emissions in the relevant region must have also been constant to within a few tens of per cent. The relevant region (based on model estimates for Vostok) is the Indian Ocean sector, with the largest influence between 55° and 60°S ¹⁰.

Marine proxies have suggested that export production south of the present-day Antarctic polar front (APF) was lower in the LGM than in the Holocene³¹. Our result implies that organisms such as *Phaeocystis*, which leave no fossil record, but which are major DMS producers, also had no increased production south of the APF. In fact, production by these organisms remained approximately constant, despite the large changes in Antarctic environmental conditions and the decrease in overall export production. Our measurements are probably less sensitive to changes in DMS production further north. Because DMS emissions south of the APF apparently remained constant to within a few tens of per cent throughout the last 740 kyr, despite very large climate changes, the sulphur cycle does not appear to have responded to climate change with either a positive or negative feedback.

Iron flux and South American climate

The nssCa flux, as also does the Fe flux, shows extreme variations, being about a factor of ten higher in glacial maxima than in interglacial periods (Fig. 2). Both metals are assumed to be representative of terrestrial material, and their concentration profiles look similar to that of insoluble dust mass⁶. The main origin of dust reaching these sites is South America, specifically Patagonia³². Model studies^{28,29}, and previous studies of dust particle size for the more recent part of the record³³, suggest that changes in the transport strength and atmospheric residence time between South America and Dome C were small, despite the changes in precipitation scavenging and atmospheric circulation that must have occurred. Assuming this applies to the whole record then the order of magnitude changes in dust flux seen at Dome C must result primarily from changes in the source region.

Such changes in South America could result from (1) changes in the strength of the Patagonian source due to changes in temperature, moisture and vegetation, (2) changes in uplift of source material due to changes in wind strength, (3) an increased Patagonian source of fine glacial material due to varying glacier coverage, or (4) the addition of new source areas on exposed continental shelf as sea level changed. The latest evidence³⁴ suggests that geochemical measurements cannot yet distinguish between a Patagonian and continental shelf source.

It is very likely that climate conditions in Patagonia did change during glacial periods. A recent model study suggested that, south of 50°S , the westerlies were intensified at the LGM, and that this was coupled to a reduction in moisture transport to Patagonia from the Atlantic (giving drier conditions)³⁵. If the main dust source is well south, then it should be enhanced by such changes. Major glaciations of Patagonia occurred during cold periods in the ice-core record³⁶, although it is not obvious how this would affect the dust source. Unfortunately, there is no good time-resolved record of any of these factors that we can use to test their influence on the Antarctic dust record.

It has been argued that the detailed timing of sea-level rise is inconsistent with the timing of the main changes in dust flux during

the LGM–Holocene transition⁵. Most of the area of continental shelf south of 45°S that would have been exposed during the LGM would have been re-flooded early in the deglaciation³⁷ (see Supplementary Information). Nonetheless, the nssCa flux had already halved from its LGM maximum by 17 kyr BP, before any significant change in sea level³⁸ or exposed area, and the nssCa flux reduction clearly leads the flooding of the shelf. Flooding of the shelf cannot be the main cause of reduced nssCa flux, although the changing exposure of continental shelf could play some part. There is a general correspondence between particularly high marine benthic oxygen isotope values³⁹, representing—at least in part—changes in sea level, and high nssCa fluxes (Fig. 2); however, there is also a strong correspondence between nssCa flux and other parameters such as Antarctic δD .

We therefore conclude that the nssCa record probably mainly reflects changes in atmospheric circulation influencing Patagonian wind strength and aridity, although changes in glaciation might also have an unspecified role. The increased exposure of continental shelf may amplify the effect when sea level is particularly low. NssCa flux, in some way a record of Patagonian conditions, shows a very close, but nonlinear, link between Patagonian conditions and Antarctic climate throughout the period. NssCa flux (and similarly Fe flux) increases significantly only below a threshold (about -410‰ at Dome C) in Antarctic deuterium (Fig. 3). Fluxes are similar in the first and second halves of the record for a given deuterium value (Fig. 3); slightly higher glacial values earlier in the record rest heavily on a few high flux values in MIS12 and MIS16. There seems to be no significant change in the climate–dust feedback at 440 kyr. Finally, although Australia could also be a significant source of dust, models suggest that Patagonia should also be the main source of atmospheric iron to most of the Southern Ocean⁴⁰, and thus our record should

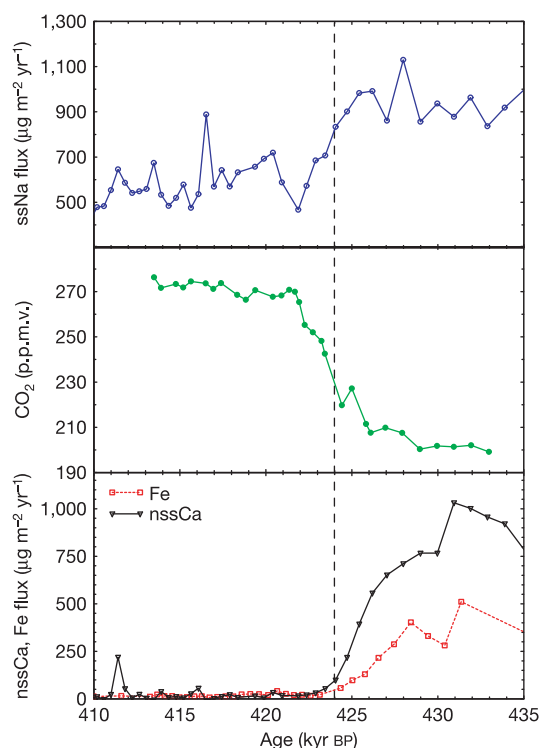


Figure 4 | Chemistry and CO_2 across Termination V. Chemical fluxes (averaged over 1.1 m depth increments, equivalent to a few hundred years at this depth, except for Fe, which consists of spot values at irregular intervals), and CO_2 concentration⁶, across Termination V, between MIS12 and MIS11. Uncertainty in the alignment of the timescales for the CO_2 and chemical records is caused by the calculation of the gas-age/ice-age difference⁶, and could be several centuries. The vertical dashed line indicates 424 kyr BP (see text).

serve both as a continuous profile of changes in Patagonian climate, and as a suitable input for assessing the influence of Fe fertilization on atmospheric CO₂ concentration.

Phasing of changes at climatic transitions

Although many aspects of climate throughout the 740-kyr record change together at each termination and inception, this is not the case, at least at millennial resolution, for sea salt (representing sea ice) and terrestrial material (representing Patagonian conditions). It was previously noted^{5,9,41} that, at Termination I, the main change in nssCa flux occurred in the first half of the transition (18–15 kyr), while the main change in Na flux occurred later (14–12 kyr).

The new record shows that this is a general rule at terminations, as illustrated in detail for Termination V (Fig. 4). In this termination, the main change in terrestrial material flux (at 428–424 kyr) precedes that in sea salt (425–422 kyr), while the CO₂ change occurs mainly between 426 and 422 kyr BP. Deuterium increases between 428 and 422 kyr. The same pattern is seen also at the earlier changes, where there is a lower amplitude in δD (the phasing between CO₂ (ref. 42) and our parameters has yet to be determined at these transitions): at the transition from MIS16 to MIS15, for example, the reduction in nssCa occurs mainly between 629–625 kyr, and the reduction in ssNa occurs mainly between 625–623 kyr BP. This pattern suggests that the main changes in Patagonia generally precede those in sea-ice extent.

It also has strong implications for the possible causes of CO₂ increases during terminations, because mechanisms related to changing Fe fertilization⁴³ can only be active in the first half of the termination, unless Fe ceases to be the limiting nutrient at deposition rates only just above those seen in the Holocene. In Termination V, at 424 kyr, when the terrestrial dust change is almost complete, CO₂ has increased by <30 p.p.m.v., adding to evidence^{9,44} that this is a maximum possible impact of iron fertilization. We note that the increased productivity implied by iron fertilization would have to occur north of the APF. Mechanisms related to changing sea-ice extent²³, at least in the ocean sector represented in our record, should be active in the later stages of each transition only.

Climate and chemistry over eight cycles

Using only Antarctic ice-core chemical data, we describe the dynamics of three external parts of the climate system over a period of 740 kyr, and provide continuous data sets as inputs for climate modelling. According to our interpretation, maximum sea-ice extent in the Indian Ocean sector is tied very closely to Antarctic temperature at multimillennial timescales, but they are not synchronous at short timescales. The part of marine biogenic activity in the Southern Ocean that is related to DMS emission south of the APF seems to have been remarkably constant through the period; atmospheric S compounds had no significant role in climate feedback. Patagonian conditions changed very strongly between glacial and interglacial, leading to a large change in iron input to the Southern Ocean.

The relationship between each chemical component and Antarctic deuterium (temperature) stays remarkably constant through the last 740 kyr, with no significant change at the point where the amplitude of interglacials increases strongly (around 440 kyr). We conclude that change in internal feedbacks represented by these parameters was not responsible for the change in amplitude of interglacial climate.

METHODS

Major ions presented here (Na⁺, Ca²⁺, SO₄²⁻) were measured using IC (ion chromatography), with an estimated uncertainty on individual measurements of better than 5% (but rather higher at the lowest concentrations seen in interglacial periods)⁴⁵. Samples from 0–580 m depth were cut into discrete samples from a section of the core after removal of the contaminated outer layers, melted and analysed; below 580 m, they were collected into sample vials using a melting device. In the latter case, a 3.4 × 3.4 cm strip of ice was melted onto a hotplate (either in the field or in a European cold laboratory)⁴⁶, and part of the melt from the inner part of the core was led directly to various detection devices (in a

continuous flow analysis (CFA) system), or fed into vials for later IC analysis. Of the components shown here, Na⁺ and Ca²⁺ were determined spectrometrically using the CFA system⁴⁶ as well as by IC. The results shown here are from IC, but good agreement between methods is generally found⁴⁵. The exception is for Ca²⁺ in the upper 580 m, for which the IC data may be slightly contaminated owing to possible incomplete removal of outer layers. This problem does not affect any other ions, and we found good agreement between IC and CFA Ca²⁺ data, even at the low concentrations found in interglacial periods, beyond 580 m. We have therefore used the CFA Ca²⁺ data for the upper 580 m (27 kyr). SO₄²⁻ was measured by passing some of the CFA water directly into a fast-IC device⁴⁷, as well as by conventional IC. A comparison between the two methods shows excellent agreement, but hereafter conventional IC data are used. None of the three methods described here is likely to measure the insoluble fraction of terrestrial dust. For the IC measurements, samples were collected as the cumulative melt or cut from anything between 5 cm to 1.1 m of ice, and in Fig. 1 we present averages of 2.2 m increments.

Fe was measured on 1,038 discrete samples using inductively coupled plasma sector field mass spectrometry (ICP-SFMS). Ice-core sections 5 cm long were decontaminated by three repeat washings in ultrapure water to remove 60% of the material, acidified to pH 1 with ultrapure nitric acid, and then left to stand for at least 24 h before analysis⁴⁸. Precision is around 20% for interglacial and 10% for glacial samples. The discrete samples are rather widely spaced, so that the time averages seen in Fig. 2 often consist of one or more data points representing only a few per cent of the time interval. Some of the detailed differences between nssCa and Fe arise from the fact that the nssCa data are averaged from samples covering most of the averaging interval, in contrast to the Fe data. It is for this reason that we prefer to use nssCa for most of the paper.

The timescale is EDC2 (ref. 6), and the accumulation rates used to calculate flux also derive from EDC2. We calculated ssNa and nssCa assuming a Ca/Na weight ratio of 0.038 for marine aerosols and 1.78 for the average crust⁴⁹. Although the latter figure could be highly variable between different terrestrial sources, leading to a substantial uncertainty in ssNa, previous authors^{5,50} have suggested that this is the most suitable ratio to use. If we had used crustal ratios with Fe as the terrestrial marker element, the correction would have been smaller. If the crustal source material is dominated by marine clays or other Na-rich material the correction would be larger. In any case, no reasonable ratios affect the shape of the derived time series, nor the conclusions of this paper. On average we calculate that about 90% of Na is from sea salt, although the proportion occasionally falls below 70% in glacial periods. In glacial periods, about 90% of Ca is from terrestrial sources, but the proportion can be much lower in interglacials.

nssSO₄²⁻ was first derived by subtracting the sea-salt part, traditionally calculated by using ssNa, along with the weight ratio of SO₄²⁻/Na in sea water (0.25). However, if the main source of sea salt is actually the sea-ice surface, then this source is depleted in sulphate compared to sea water, with a ratio close to 0.1 (refs 17, 18). We present nssSO₄²⁻ fluxes calculated with the latter method—again, our choice does not significantly affect the results of this paper.

Received 5 September 2005; accepted 26 January 2006.

- Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).
- Legrand, M. *et al.* Ice-core record of oceanic emissions of dimethylsulphide during the last climate cycle. *Nature* **350**, 144–146 (1991).
- Wagnon, P., Delmas, R. J. & Legrand, M. Loss of volatile acid species from upper firn layers at Vostok, Antarctica. *J. Geophys. Res.* **104**, 3423–3431 (1999).
- Weller, R. *et al.* Post depositional losses of methane sulphonate, nitrate, and chloride at the EPICA deep-drilling site in Dronning Maud Land, Antarctica. *J. Geophys. Res.* **109**, 10.1029/2003JD004189 (2004).
- Röthlisberger, R. *et al.* Dust and sea salt variability in central East Antarctica (Dome C) over the last 45 kyr and its implications for southern high-latitude climate. *Geophys. Res. Lett.* **29**, 1963, doi:10.1029/2002GL015186 (2002).
- EPICA Community Members, Eight glacial cycles from an Antarctic ice core. *Nature* **429**, 623–628 (2004).
- Legrand, M. Chemistry of Antarctic snow and ice. *J. Phys.* **48**, C1/77–86 (1987).
- Martin, J. H. & Fitzwater, S. E. Iron deficiency limits phytoplankton growth in the north-east Pacific subarctic. *Nature* **331**, 341–343 (1988).
- Röthlisberger, R. *et al.* Ice core evidence for the extent of past atmospheric CO₂ change due to iron fertilisation. *Geophys. Res. Lett.* **31**, L16207, doi:10.1029/2004GL020338 (2004).
- Cosme, E., Hourdin, F., Genthon, C. & Martinier, P. Origin of dimethylsulphide, non-sea-salt sulfate, and methanesulfonic acid in eastern Antarctica. *J. Geophys. Res.* **110**, D03302, doi:10.1029/2004JD004881 (2005).
- Legrand, M. in *Ice Core Studies of Global Biogeochemical Cycles* (ed. Delmas, R. J.) 91–119 (Springer, Berlin, 1995).

12. Castellano, E. *et al.* Volcanic eruption frequency over the last 45 ky as recorded in Epica-Dome C ice core (East Antarctica) and its relationship with climatic changes. *Glob. Planet. Change* **42**, 195–205 (2004).
13. Castellano, E. *et al.* Holocene volcanic history as recorded in the sulfate stratigraphy of the European Project for Ice Coring in Antarctica Dome C (EDC96) ice core. *J. Geophys. Res.* **110**, D06114, doi:10.1029/2004JD005259 (2005).
14. Delmas, R. J., Wagnon, P., Goto-Azuma, K., Kamiyama, K. & Watanabe, O. Evidence for the loss of snow-deposited MSA to the interstitial gaseous phase in central Antarctic firn. *Tellus B* **55**, 71–79 (2003).
15. Rothlisberger, R., Hutterli, M. A., Sommer, S., Wolff, E. W. & Mulvaney, R. Factors controlling nitrate in ice cores: evidence from the Dome C deep ice core. *J. Geophys. Res.* **105**, 20565–20572 (2000).
16. Wolff, E. W., Rankin, A. M. & Rothlisberger, R. An ice core indicator of Antarctic sea ice production? *Geophys. Res. Lett.* **30**, 2158, doi:10.1029/2003GL018454 (2003).
17. Wagenbach, D. *et al.* Seasalt aerosol in coastal Antarctic regions. *J. Geophys. Res.* **103**, 10961–10974 (1998).
18. Rankin, A. M., Wolff, E. W. & Martin, S. Frost flowers—implications for tropospheric chemistry and ice core interpretation. *J. Geophys. Res.* **107**, 4683, doi:10.1029/2002JD002492 (2002).
19. Rankin, A. M., Wolff, E. W. & Mulvaney, R. A reinterpretation of sea salt records in Greenland and Antarctic ice cores. *Ann. Glaciol.* **39**, 276–282 (2005).
20. Reader, M. C. & McFarlane, N. Sea-salt aerosol distribution during the Last Glacial Maximum and its implications for mineral dust. *J. Geophys. Res.* **108**, 4253, doi:10.1029/2002JD002063 (2003).
21. Hara, K. *et al.* Chemistry of sea-salt particles and inorganic halogen species in Antarctic regions: Compositional differences between coastal and inland stations. *J. Geophys. Res.* **109**, D20208, doi:10.1029/2004JD004713 (2004).
22. Gersonde, R., Crosta, X., Abellmann, A. & Armand, L. Sea-surface temperature and sea ice distribution of the Southern Ocean at the EPILOG Last Glacial Maximum—a circum-Antarctic view based on siliceous microfossil records. *Quat. Sci. Rev.* **24**, 869–896 (2005).
23. Stephens, B. B. & Keeling, R. F. The influence of Antarctic sea ice on glacial–interglacial CO₂ variations. *Nature* **404**, 171–174 (2000).
24. Charlson, R. J., Lovelock, J. E., Andreae, M. O. & Warren, S. G. Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature* **326**, 655–661 (1987).
25. Kirchner, J. W. The Gaia hypothesis: Conjectures and refutations. *Clim. Change* **58**, 21–45 (2003).
26. Lovelock, J. A geophysicist's thoughts on the natural sulphur cycle. *Phil. Trans. R. Soc. Lond. B* **352**, 143–147 (1997).
27. Watanabe, O. *et al.* General tendencies of stable isotopes and major chemical constituents of the Dome Fuji deep ice core. *Mem. Nat. Inst. Polar Res. (Spec. Iss.)* **57**, 1–24 (2003).
28. Krinner, G. & Genthon, C. Tropospheric transport of continental tracers towards Antarctica under varying climatic conditions. *Tellus B* **55**, 54–70 (2003).
29. Lunt, D. J. & Valdes, P. J. Dust transport to Dome C, Antarctica, at the Last Glacial Maximum and present day. *Geophys. Res. Lett.* **28**, 295–298 (2001).
30. Legrand, M. & Pasteur, E. C. Methane sulfonic acid to non sea-salt sulfate ratio in coastal Antarctic aerosol and surface snow. *J. Geophys. Res.* **103**, 10991–11006 (1998).
31. Kohfeld, K. E., Le Quere, C., Harrison, S. P. & Anderson, R. F. Role of marine biology in glacial–interglacial CO₂ cycles. *Science* **308**, 74–78 (2005).
32. Basile, I. *et al.* Patagonian origin of glacial dust deposited in East Antarctica (Vostok, Dome C) during glacial stages 2, 4 and 6. *Earth Planet. Sci. Lett.* **146**, 573–589 (1997).
33. Delmonte, B., Petit, J. R. & Maggi, V. Glacial to Holocene implications of the new 27000-year dust record from the EPICA Dome C (East Antarctica) ice core. *Clim. Dyn.* **18**, 647–660 (2002).
34. Delmonte, B. *et al.* Comparing the Epica and Vostok dust records during the last 220,000 years: stratigraphical correlation and provenance in glacial periods. *Earth Sci. Rev.* **66**, 63–87 (2004).
35. Wainer, I., Clauzet, G., Ledru, M.-P., Brady, E. & Otto-Bliesner, B. Last glacial maximum in South America: Paleoclimate proxies and model results. *Geophys. Res. Lett.* **32**, L08702, doi:10.1029/2004GL021244 (2005).
36. Kaplan, M. R., Douglass, D. C., Singer, B. S. & Caffee, M. W. Cosmogenic nuclide chronology of pre-last glacial maximum moraines at Lago Buenos Aires, 46 degrees S, Argentina. *Quat. Res.* **63**, 301–315 (2005).
37. ETOPO2 (2-Minute Gridded Global Relief Data) (www.ngdc.noaa.gov/mgg/fliers/01mgg04.html) (US Department of Commerce, National Oceanic and Atmospheric Administration, National Geophysical Data Center, Boulder, Colorado, 2001).
38. Guilderson, T. P., Burckle, L. H., Hemming, S. & Peltier, W. R. Late Pleistocene sea level variations derived from the Argentine Shelf. *Geochem. Geophys. Geosyst.* **1**, doi:10.1029/2000GC000098 (2000).
39. Lisiecki, L. E. & Raymo, M. E. A. Pliocene–Pleistocene stack of 57 globally distributed benthic delta O-18 records. *Paleoceanography* **20**, PA1003, doi:10.1029/2004PA001071 (2005).
40. Lunt, D. J. & Valdes, P. J. Dust deposition and provenance at the Last Glacial Maximum and present day. *Geophys. Res. Lett.* **29**, doi:10.1029/2002GL015656 (2002).
41. Udisti, R. *et al.* Sensitivity of chemical species to climatic changes in the last 45 kyrs as revealed by high resolution Dome C (Antarctica) ice core analysis. *Ann. Glaciol.* **39**, 457–466 (2005).
42. Siegenthaler, U. *et al.* Stable carbon cycle–climate relationship during the late Pleistocene. *Science* **310**, 1313–1317 (2005).
43. Martin, J. Glacial–interglacial CO₂ change: The iron hypothesis. *Paleoceanography* **5**, 1–13 (1990).
44. Bopp, L., Kohfeld, K. E., Le Quere, C. & Aumont, O. Dust impact on marine biota and atmospheric CO₂ during glacial periods. *Paleoceanography* **18**, 1046, doi:10.1029/2002PA000810 (2003).
45. Littot, G. C. *et al.* Comparison of analytical methods used for measuring major ions in the EPICA Dome C (Antarctica) ice core. *Ann. Glaciol.* **35**, 299–305 (2002).
46. Rothlisberger, R. *et al.* Technique for continuous high-resolution analysis of trace substances in firn and ice cores. *Environ. Sci. Technol.* **34**, 338–342 (2000).
47. Traversi, R. *et al.* High-resolution fast ion chromatography (FIC) measurements of chloride, nitrate and sulphate along the EPICA Dome C ice core. *Ann. Glaciol.* **35**, 291–298 (2002).
48. Planchon, F. A. M. *et al.* Ultrasensitive determination of heavy metals at the sub-picogram per gram level in ultraclean Antarctic snow samples by inductively coupled plasma sector field mass spectrometry. *Anal. Chim. Acta* **450**, 193–205 (2001).
49. Bowen, H. J. M. *Environmental Chemistry of the Elements* (Academic Press, London, 1979).
50. Legrand, M. R. & Delmas, R. J. Soluble impurities in four Antarctic ice cores over the last 30 000 years. *Ann. Glaciol.* **10**, 116–120 (1988).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank all the EPICA personnel who have contributed to obtaining the ice-core samples, either through logistics, drilling or core processing. This work is a contribution to EPICA, a joint European Science Foundation/European Commission (EC) scientific programme, funded by the EC (EPICA-MIS) and by national contributions from Belgium, Denmark, France, Germany, Italy, The Netherlands, Norway, Sweden, Switzerland and the UK.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.W.W. (ewwo@bas.ac.uk).

Human chromosome 11 DNA sequence and analysis including novel gene identification

Todd D. Taylor¹, Hideki Noguchi^{1†}, Yasushi Totoki¹, Atsushi Toyoda¹, Yoko Kuroki¹, Ken Dewar^{2†}, Christine Lloyd³, Takehiko Itoh⁴, Tadayuki Takeda¹, Dae-Won Kim⁵, Xinwei She⁶, Karen F. Barlow³, Toby Bloom², Elspeth Bruford⁷, Jean L. Chang², Christina A. Cuomo², Evan Eichler⁶, Michael G. FitzGerald², David B. Jaffe², Kurt LaButti², Robert Nicol², Hong-Seog Park⁵, Christopher Seaman², Carrie Sougnéz², Xiaoping Yang², Andrew R. Zimmer², Michael C. Zody², Bruce W. Birren², Chad Nusbaum², Asao Fujiyama^{1,8}, Masahira Hattori^{1,9}, Jane Rogers³, Eric S. Lander² & Yoshiyuki Sakaki¹

Chromosome 11, although average in size, is one of the most gene- and disease-rich chromosomes in the human genome. Initial gene annotation indicates an average gene density of 11.6 genes per megabase, including 1,524 protein-coding genes, some of which were identified using novel methods, and 765 pseudogenes. One-quarter of the protein-coding genes shows overlap with other genes. Of the 856 olfactory receptor genes in the human genome, more than 40% are located in 28 single- and multi-gene clusters along this chromosome. Out of the 171 disorders currently attributed to the chromosome, 86 remain for which the underlying molecular basis is not yet known, including several mendelian traits, cancer and susceptibility loci. The high-quality data presented here—nearly 134.5 million base pairs representing 99.8% coverage of the euchromatic sequence—provide scientists with a solid foundation for understanding the genetic basis of these disorders and other biological phenomena.

Human chromosome 11 (HSA11), which represents approximately 4.4% of the human genome^{1,2}, has had a significant role in the history of molecular genetics, beginning long before its complete sequencing was undertaken. The haemoglobin beta gene, encoding one of the best-studied proteins, was one of the first genes mapped to the human genome (11p15.5) and was the first protein to have its crystal structure solved³. It is also the cause of sickle cell anaemia, the first human genetic disease for which a molecular basis was demonstrated⁴. Three megabases (Mb) distal lies the insulin gene, encoding the first fully-sequenced protein⁵, and the intensely studied imprinting region responsible for Beckwith–Wiedemann syndrome⁶. The physical map, high-quality finished sequence and gene catalogue presented here are but the latest landmark in an effort to understand the unique characteristics and functions of this chromosome.

The clone map and finished sequence

Chromosome 11 was sequenced using a clone-by-clone shotgun sequencing approach. The sequence is in eight finished contigs (Supplementary Tables S1–S3), the largest being 49.6 Mb, with seven gaps remaining, including one at 11p-tel (~50 kilobases (kb)), one heterochromatic gap (207 kb) near 11p-cen and five small internal clone gaps (totalling ~64.5 kb). Where possible, all of the gaps were size-estimated by fibre-fluorescence *in situ* hybridization (FISH) analysis. On 11q, we reached both the telomeric

repeats and the centromeric alpha satellite repeats, and higher-order repeat structure was observed in clone AC126345 at 11p-cen. To ensure production of the most reliable data, sequence quality control checks were performed both internally (Supplementary Table S4) and externally⁷. In total, we finished 131,130,853 base pairs (bp) and estimate the total size of the chromosome, including the gaps and centromere, to be approximately 134.5 Mb (May 2004, NCBI build 35). The coverage of the euchromatic portion of the chromosome is an estimated 99.8%. Of the finished sequence, 60% was generated by RIKEN Genomic Sciences Center, 36% by the Broad Institute of MIT and Harvard, 3% by the Wellcome Trust Sanger Institute, and 1% by the Washington University School of Medicine Genome Sequencing Center.

The chromosome landscape

Figure 1 shows an overview of the chromosome 11 landscape. HSA11 is very gene rich and there are many clustered gene families located on the chromosome. According to a recent survey of the Ensembl genome browser⁸, HSA11 contains the fourth highest number of genes in the human genome, after human chromosomes 1, 2 (ref. 9) and 19 (ref. 10), respectively. These data show 10.6 protein-coding genes per Mb on HSA11, as compared to the genome-wide average of 7.3. In fact, manual annotation of the chromosome identifies a slightly higher gene density of 11.6 genes per Mb, with genes spaced

¹RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. ²Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141, USA. ³The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ⁴Mitsubishi Research Institute, Inc., 2-3-6 Otemachi, Chiyoda-ku, Tokyo 100-8141, Japan. ⁵Korea Research Institute of Bioscience & Biotechnology, 52 Oun-dong, Yusong-gu, Daejeon 305-333, South Korea. ⁶University of Washington, Genome Sciences, HSB K336B, Box 357730, Seattle, Washington 98195, USA. ⁷HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK. ⁸National Institute of Informatics, Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo 101-8430, Japan. ⁹Kitasato Institute for Life Sciences, Kitasato University 1-15-1, Kitasato, Sagami-hara, Kanagawa 228-8555, Japan. [†]Present addresses: University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-0882, Japan (H.N.); McGill University and Genome Quebec Innovation Centre, Montreal, Quebec H3A 1A4, Canada (K.D.).

an average of 86 kb apart. Both the repeat density (47.98%; Supplementary Table S5 and Supplementary Information) and G+C content (41.57%) are close to genome-wide averages. Table 1 lists various features of the chromosome.

The finished sequence of HSA11 shows strong concordance with existing physical and genetic maps. All sequence-tagged sites from the Génethon microsatellite-based genetic map¹¹, the deCODE map¹² and the Marshfield genetic maps¹³ are present in the HSA11 sequence (Supplemental Methods). We compared recombination rates in the deCODE female, male and sex-averaged meiotic maps (which average 1.53, 0.85 and 1.19 cM per Mb, respectively) with the physical distance as determined from the sequence assembly (Supplementary Fig. S1). Recombination statistics for HSA11 are similar to other human chromosomes, showing a relatively linear relationship between recombination rate and physical distance.

Gene catalogue

We annotated a total of 2,347 gene loci consisting of 1,524 potentially active protein-coding genes, 765 pseudogenes and 58 RNAs (Supplementary Table S6 and Supplementary Methods). The 1,524 protein-coding genes comprise 1,195 known genes (including 166 olfactory receptor genes), 104 novel coding sequences (CDSs), 221 novel transcripts and four putative genes. Some of these genes were identified by our *ab initio* gene prediction program DIGIT¹⁴, as described below. The 765 pseudogenes include at least three unprocessed pseudogenes and 203 olfactory receptor pseudogenes. In total, we annotated 230 previously unknown genes (that is, no RefSeq or Ensembl location, Supplementary Methods) consisting of 48 novel CDSs, 178 novel transcripts and four putative genes. These novel genes are scattered throughout the chromosome, with many located in potential disease candidate regions.

There are 296 single-exon genes, of which 168 belong to the olfactory receptor gene family. The remaining 1,228 multi-exon genes (80.53%) have an average of 9.39 exons per gene. In addition to the olfactory receptor gene clusters described later, we identified 142 genes in 37 clusters that belong to gene families with at least two members on HSA11 (Supplementary Table S7).

Co-transcribed or read-through genes do not appear to be a very common phenomenon in the human genome, but this could be due to the current lack of uniform genome-wide gene annotation. We found 12 cases on HSA11 (Supplementary Table S8), which are each supported by just one messenger RNA (and in a few cases by expressed sequence tags (ESTs)). Besides these examples, we found only a few other examples on chromosomes 17 and 22 (ref. 15) (Supplementary Methods). Of these, only two were found that probably result in a protein fusion product, *TRIM6-TRIM34* and *BSCL2-HNRPUL2*. Whether or not these read-through transcripts should be considered as alternative transcripts or separate genes, with functions different from the two genes they connect, remains to be investigated. Because the supporting evidence for such read-through transcripts is usually minimal, additional experiments should first be carried out to determine whether or not they are real, or just represent cellular mistakes or artefacts.

For the protein-coding genes, we attempted to identify all possible splice variants using currently available mRNA data (and, in a few cases, EST information). We found that 805 (52.8%) of the genes have at least two or more variants, consisting of 738 known genes, 36 novel CDSs, 30 novel transcripts and one putative gene. The genes with at least two variants have an average of 3.73 variants per locus. The *CTNND1* gene showed the largest number of variants with 28. In total, we identified 3,723 variants for the 1,524 expressed genes. Of these nearly 4,000 splice variants, there are many instances where the transcripts splice correctly but do not have definitive or long (>100 amino acids) open reading frames and may be examples of incompletely spliced RNAs, incorrectly spliced RNAs or non-coding RNAs.

We explored whether there was any correlation between the

presence of a CpG island and the number of variant transcripts for a gene (Supplementary Table S9). Interestingly, we found a significant correlation ($\chi^2 = 224.29$, $P < 0.0001$, 6 degrees of freedom). Out of the 894 genes with CpG islands, 650 (70%) have two or more variants. By contrast, of the 626 genes with no CpG islands, only 154 (24.6%) have two or more variants.

Olfactory receptor genes

Olfactory receptor genes comprise the largest multi-gene family in metazoans. All human chromosomes except HSA20 (ref. 16) and HSAY (ref. 17) contain olfactory receptor genes, but HSA11 is by far the richest. In human there are 856 olfactory receptor genes, 369 (43%) of which are located on HSA11 (ref. 18). These are mostly single-exon genes, with an average length of about 1 kb. Of the 369 loci on HSA11, 166 (45%) are protein-coding and the other 203 (55%) are pseudogenes; this is close to the genome-wide average (47% versus 53%). All but 10 of the olfactory receptor genes on HSA11 lie within 18 clusters, separated by at least 100 kb (Figs 1 and 2; see also Supplementary Table S10). The largest cluster contains 97 genes over a range of 1.5 Mb. The average distance between genes within a cluster is about 17 kb. The olfactory receptor genes on HSA11 are classified into 13 different families (having >40% protein identity), containing from as few as one to as many as 81 members. The olfactory receptor regions on HSA11 generally are rich in L1 repeats, poor in Alu repeats, CpG islands (Supplementary Table S11 and Supplementary Information) and predicted transcription starts (based on the Eponine program¹⁹), and have a G+C content of 40% or lower. Functional olfactory receptor genes are evenly distributed within the clusters.

Olfactory receptor genes are roughly classified into two classes: I and II. Class I olfactory receptor genes are known as fish-like olfactory receptors and are believed to be receptors for water-soluble ligands. They have expanded in mammalian lineages (Fig. 2) and many belong to one large cluster in mammalian genomes. In the human genome, all of the class I olfactory receptor genes are found in three closely spaced clusters on the subtelomeric region of 11p, from

Figure 1 | Chromosome 11 landscape. The following sections appear in order from top to bottom. (1) Cytogenetic banding pattern. (2) Non-olfactory receptor (OR) expressed genes. Genes are colour coded according to category type (known genes, red; novel CDS, green; novel transcript, blue; putative, orange) and are shown according to their orientation on the chromosome (upper, forward orientation; lower, reverse orientation: note that this applies to sections 2–6, which means that these sections appear twice (with sections 7–9 sandwiched between them), once for forward and once for reverse orientation). (3) Non-olfactory receptor clustered gene families. (4) Non-olfactory receptor pseudogenes. (5) Miscellaneous RNAs including tRNAs, microRNAs and snoRNAs. (6) Eponine TSS predicted transcription start sites. (7) CpG islands. (8) Exofish evolutionarily conserved regions (ECRs). (9) Gene deserts larger than 650 kb. (10) Olfactory receptor genes. The olfactory receptor genes are colour coded (connecting line, see key on left) according to family and whether a coding sequence (blue symbol name) or pseudogene (grey symbol name). Upper, forward orientation on chromosome; lower, reverse orientation on chromosome. (11) G+C content. This was calculated using a sliding window of 100 kb. (12) Short interspersed element (SINE)/Alu (blue) and long interspersed element (LINE)/L1 (red) repeat densities. This was also calculated using a sliding window of 100 kb. (13) Interspersed repeat elements as reported by RepeatMasker (SINEs, LINEs, long terminal repeat (LTR) elements, DNA elements, unclassified repeats, simple repeats, low-complexity regions, RNAs and satellite repeats). Also included in this region are tandem repeats as predicted by the Tandem Repeat Finder program. (14) Interchromosomal duplications (red) and intrachromosomal duplications (blue). (15) Gene disorders (mapped only). Forty-one disorders that have been mapped to HSA11 are shown as coloured horizontal bars, which represent the possible genetic location of the disorder according to OMIM (Online Mendelian Inheritance in Man). Clone gaps (including the centromere) are indicated by the vertical grey blocks.

Table 1 | Chromosome 11 sequence features

Chromosome property	Value
Chromosome length (bp)	134,452,3843
Finished sequence length (bp)	131,130,853
Total gap length including 3 Mb centromere (bp)	3,321,531
Euchromatic region coverage (%)	99.76
Protein-coding gene loci	1,524
Known	1,195
Novel CDS	104
Novel transcripts	221
Putative	4
Pseudogenes	765
tRNAs (all)	19
tRNA pseudogenes	2
MicroRNAs	12
snoRNAs	27
Gene coverage (bp)	61,581,901 (46.96%)
Exon coverage (bp)	3,543,101 (2.70%)
Known gene mean size (bp)	45,007
CpG islands	1,369
Genes with 5' CpG islands	806
Gene deserts (>650 kb)	19
Gene desert coverage (bp)	23,378,900 (17.83%)

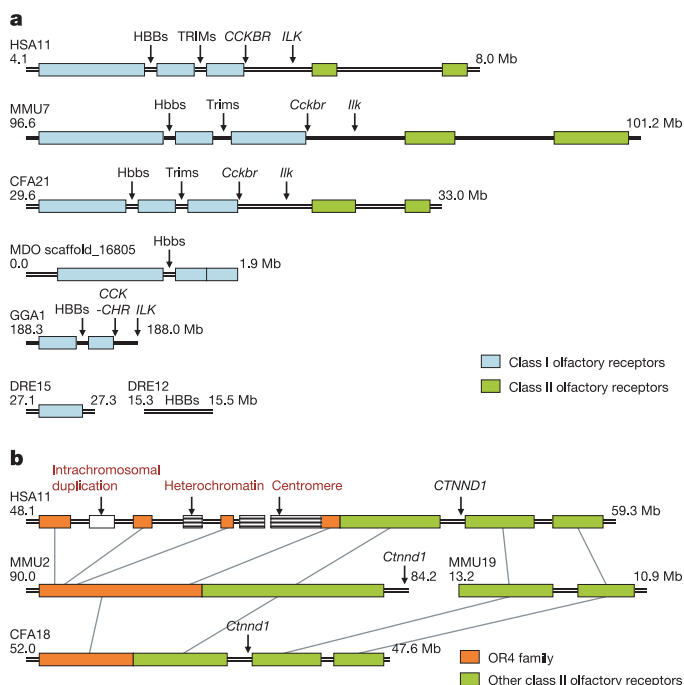
snoRNAs, small nucleolar RNAs.

4.1 to 6.2 Mb. Approximately 50% (54 out of 103) of these genes are intact in human, which is close to the genome-wide average for all olfactory receptors. The class I region is interrupted by a few genes including the beta-globin gene cluster and a TRIM gene cluster.

The most significant class II cluster is located around the centromere of HSA11. The corresponding clusters for the mouse²⁰, rat²¹ and dog²² genomes are also the largest ones, and are comprised of many families of class II olfactory receptor genes. Notably, the human cluster has a significantly different structure from that of other mammals: the human cluster is divided by insertion of the centromere, a heterochromatic region and an intrachromosomal duplication. Despite the structural changes, and although some members of the cluster are on different chromosomes in rodents, analysis of conserved order of orthologous sequences suggests that they belonged to one large cluster in their last common ancestral genome.

Identification of weakly expressed novel genes

As mentioned above, we applied DIGIT¹⁴ (Methods), an *ab initio* gene-finder, to HSA11. The program predicted 65 novel protein-coding gene loci, with an average open-reading length of 366 amino acids, the genomic regions of which did not overlap at the time with human mRNAs from GenBank. We found experimental support for 34 (52%) of these predictions, based on reverse transcription polymerase chain reaction (RT-PCR) experiments that show evidence of the predicted splice junctions (including four full CDSs) (Supplementary Tables S12 and S13). Most (26 out of 34) of these genes appear to be expressed at an extremely low level (detectable only by nested PCR), which might explain why they had not been previously detected by any high-throughput EST or full-length cDNA sequencing strategy. Of the 34 genes with experimental support (Supplementary Table S14), 12 were identified only through this method because they have no orthologous sequence in any currently available genome (six also have no related human sequence, whereas six have related human sequence). Eight of the 34 genes may simply be extensions of nearby human genes. The remaining 14 genes either have orthologous sequence in other species or are highly similar to known human genes. Many of the 34 genes are predicted by the InterProScan²³ program to contain a functional domain (Supplementary Table S15). Further experimental evidence to support the expression of these genes and to identify their full-length structures is necessary. In order to obtain a more complete catalogue of all protein-coding genes in the human genome, this type of analysis should ideally be extended to include all chromosomes, especially as

**Figure 2 | Conservation and expansion of olfactory receptor clusters.**

a, Structure of the class I olfactory receptor genes and neighbouring class II genes (HSA11 4.1–8.0 Mb). The basic structure of the class I cluster is highly conserved. The beta-globin gene cluster was inserted after the divergence of fish and other vertebrates. The TRIM cluster was inserted in the placental mammalian lineage. **b**, The olfactory receptor 4 (OR4) superfamily cluster (HSA11 48.1–55.2 Mb) was divided into four pieces by insertion of the centromere, a heterochromatic region and an intrachromosomal duplication. Overall, the olfactory receptor gene clusters in mouse were expanded. HSA11, human chromosome 11; MMU2, MMU7 and MMU19, mouse chromosomes 2, 7 and 19; CFA18 and CFA21, dog chromosomes 18 and 21; MDO, opossum; GGA1, chicken chromosome 1; DRE12 and DRE15, zebrafish chromosomes 12 and 15.

some genes were only identified by this ‘*ab initio* plus experimental verification’ approach.

Conclusions

This work describes just a few of the interesting features of human chromosome 11. Notably, the chromosome is very rich in genes overall and disease genes in general (Supplementary Fig. S2). It contains many clustered gene families, the most significant being 369 members of the olfactory receptor gene family. Many medically important loci are associated with chromosome 11 for which the genetic cause of the disorder has yet to be elucidated (Supplementary Table S16). This includes various cancers, susceptibility genes and loci implicated in behavioural and psychiatric disease variation.

Some findings that stand out in our analysis include a significant correlation between the presence of CpG islands and the number of splice variants, a large number of overlapping genes (Supplementary Information) and genes sharing CpG islands, and genes that were only initially identified through *ab initio* methods. Although these phenomena may not necessarily be specific to chromosome 11, they do emphasize the need for further uniform analyses and annotation across the entire human genome. With the availability of the high-quality human genomic sequence as presented here, scientists have a solid foundation for identifying and understanding all of the genes and functional elements it holds.

METHODS

Construction of the chromosome 11 large insert libraries. We prepared chromosome-specific bacterial artificial chromosome (BAC; CMB9) and fosmid

(CMF9) libraries by using flow-sorted chromosomal DNA derived from human chromosomes 9–12 (these chromosomes cannot be separated by flow cytometry due to their similar size). For construction of the CMB9 library, sorted DNA derived from cultured lymphoblastoid cells was partially digested by *SacI* and the fragments were ligated into the pKS145 vector. Transformation was carried out by electroporation into *Escherichia coli* DH10B. The CMF9 library was prepared according to previously described methods²⁴. We screened these two libraries, the RPCI-11 whole-genome BAC library, and a few other BAC and P1-derived artificial chromosome (PAC) whole-genome libraries (Supplementary Table S2). The chromosome-specific libraries proved especially useful during the gap-filling stage of the project and for identifying clones near the complex centromeric and telomeric regions.

Clone path construction. Initial seed clones were selected by using the restriction enzyme digest fingerprint data of WUGSC and MIT for HSA11p, and by screening the RPCI-11 BAC library with evenly spaced markers taken from a highly-integrated STS map of the whole human genome for HSA11q. These approaches allowed us to construct quickly a tiling path across most of the chromosome. The remaining gaps were filled by walking from clone end sequences and by re-screening of the clone libraries. The chromosomal locations for some clones in the minimum tiling path were confirmed by FISH analysis, and the lengths of the clone gaps were estimated by fibre-FISH according to previous methods²⁵. The procedures for large-insert clone sequencing are described in Supplementary Methods.

Prediction of novel human genes. For exhaustive and efficient rare gene prediction we used DIGIT¹⁴, an *ab initio* gene-finder that finds genes by combining gene predictions from multiple *ab initio* gene-finders such as FGENESH²⁶, GENSCAN²⁷ and HMMgene²⁸. The reason we used DIGIT is that *ab initio* gene-finders, which do not use sequence similarity, have the potential for exhaustive rare gene prediction. The most remarkable feature of *ab initio* gene-finders is their high sensitivity, especially at the nucleotide level. Conversely, *ab initio* gene-finders also predict many false-positive genes. DIGIT successfully discards many false-positive exons predicted by the individual gene-finders and yields remarkable improvements in specificity without lowering sensitivity as compared with the best accuracies achieved by any single gene-finder. For experimental verification of the candidate genes, RT-PCR was performed using primer sets designed from the predicted exon sequences with a single-strand cDNA library prepared from various human tissues. If, in the first round of RT-PCR, a product could not be detected, a second round of PCR using nested PCR primer sets with the diluted RT-PCR products was conducted. When a PCR product was amplified, sequence analysis was used to confirm that the cDNA fragment was located at the predicted genomic location.

Received 17 October 2005; accepted 7 February 2006.

- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
- Bragg, W. L. & Perutz, M. F. The structure of haemoglobin. *Proc. R. Soc. Lond. A* **213**, 425–435 (1952).
- Ingram, V. M. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature* **178**, 792–794 (1956).
- Sanger, F. & Tuppy, H. The amino-acid sequence in the phenylalanyl chain of insulin 1. The identification of lower peptides from partial hydrolysates. *Biochem. J.* **49**, 463–481 (1951).
- Wiedemann, H.-R. Complexe malformatif familial avec hernie ombilicale et macroglossie – un ‘syndrome nouveau’? *J. Genet. Hum.* **13**, 223–232 (1964).
- Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429**, 365–368 (2004).
- Hubbard, T. *et al.* Ensembl 2005. *Nucleic Acids Res.* **33**, D447–D453 (2005).
- Hillier, L. W. *et al.* Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434**, 724–731 (2005).
- Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529–535 (2004).
- Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
- Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
- Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
- Yada, T., Takagi, T., Totoki, Y., Sakaki, Y. & Takaeda, Y. DIGIT: a novel gene finding program by combining gene-finders. *Pac. Symp. Biocomput.* **2003**, 375–387 (2003).
- Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
- Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).
- Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
- Olender, T., Feldmesser, E., Atarot, T., Eisenstein, M. & Lancet, D. The olfactory receptor universe – from whole genome analysis to structure and evolution. *Genet. Mol. Res.* **3**, 545–553 (2004).
- Down, T. A. & Hubbard, T. J. P. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12**, 458–461 (2002).
- Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
- Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
- Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
- Park, H.-S. *et al.* Newly identified repeat sequences, derived from human chromosome 21qter, are also localized in the subtelomeric region of particular chromosomes and 2q13, and are conserved in the chimpanzee genome. *FEBS Lett.* **475**, 167–169 (2000).
- Suto, Y., Tokunaga, K., Watanabe, Y. & Hirai, M. Visual demonstration of the organization of the human complement C4 and 21-hydroxylase genes by high-resolution fluorescence *in situ* hybridization. *Genomics* **33**, 321–324 (1996).
- Salamov, A. & Solovyev, V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
- Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
- Krogh, A. Two methods for improving performance of a HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 179–186 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Thanks to the staff, past and present, at RIKEN Genomic Sciences Center and the Broad Institute. We also acknowledge Y. Arai and M. Ohki (mapping), M. Hirai, Y. Suto and Y. Kanoh (fibre-FISH analysis technical support), C. Kawagoe and T. Katayama (computational data management), R. Baertsch and J. Mudge (annotation), V. Heyningen (historical insights), K. Linblad-Toh (preliminary assembly of the *Monodelphis domestica* genome), and the HUGO Gene Nomenclature Committee: S. Povey (chair), T. A. Eyre, V. K. Khodiyar, R. C. Lovering, K. M. B. Sneddon, T. P. Sneddon, C. C. Talbot Jr and M. W. Wright (assignment of official gene symbols). The zebrafish sequence data (assembly Zv4) were produced by the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/D_rerio/wgs.shtml). The authors also acknowledge the Ministry of Education, Culture, Sports, Science and Technology (Japan), the National Human Genome Research Institute (USA) and the Wellcome Trust Sanger Institute (UK) for funding this work.

Author Information The entire sequence for the chromosome is deposited in DDBJ/EMBL/GenBank under accession numbers NT_035113, NT_009237, NT_035158, NT_033903, NT_078088, NT_033927, NT_008984 and NT_033899. Accession numbers for individual clones and genes identified in this study can be found in Supplementary Information. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.D.T. (taylor@gsc.riken.jp) or Y.S. (sakaki@gsc.riken.jp).

Significant primordial star formation at redshifts $z \approx 3-4$

Raul Jimenez¹ & Zoltan Haiman²

Four recent observational results have challenged our understanding of high-redshift galaxies, as they require the presence of far more ultraviolet photons than should be emitted by normal stellar populations. First, there is significant ultraviolet emission¹ from Lyman break galaxies (LBGs) at wavelengths shorter than 912 Å. Second, there is strong Lyman α emission² from extended ‘blobs’ with little or no associated apparent ionizing continuum. Third, there is a population of galaxies with unusually strong Lyman α emission lines³. And fourth, there is a strong He II (1,640 Å) emission line⁴ in a composite of LBGs. The proposed explanations for the first three observations are internally inconsistent, and the fourth puzzle has remained hitherto unexplained. Here we show that all four problems are resolved simultaneously if 10–30 per cent of the stars in many galaxies at $z \approx 3-4$ are mainly primordial—unenriched by elements heavier than helium (‘metals’). Most models of hierarchical galaxy formation assume efficient intragalactic metal mixing, and therefore do not predict^{5–9} metal-free star formation at redshifts significantly below $z \approx 5$. Our results imply that micromixing of metals within galaxies is inefficient on an approximately gigayear timescale, a conclusion that can be verified with higher-resolution simulations, and future observations of the He II emission line.

The continuum emission from LBGs can be explained by adding the continuous formation of massive stars to the standard initial mass function (IMF). In the absence of such a bias in the IMF, the continuum flux emerging just below 912 Å will be dominated by low-mass stars and will be significantly fainter than observed. The other three puzzles require a substantial increase in the total number of energetic ionizing photons, and we find that this cannot be resolved by a bias of the IMF alone. To increase the number of ultraviolet photons further the temperature in the atmospheres of the stars must be increased. This is the case for metal-free (‘population III’) stars, which have significantly reduced opacity owing to the lack of elements^{10,11} C, N and O.

Lyman-continuum flux has been detected¹ in the composite spectrum of 29 LBGs at a mean redshift of $z \approx 3.4$. Although the authors of ref. 1 warn that sky background subtraction at the low observed flux level is a concern, we consider that, when taken at face value, their result implies that a surprisingly large fraction of ionizing photons can typically escape from these galaxies. These authors found $f_{\text{esc}} \approx 0.5$, where f_{esc} is defined as the ratio of 900 Å photons that escape the galaxy and the 1,500 Å photons that escape the galaxy (the wavelengths are quoted in the rest-frame of the LBGs). Using more detailed fits of the observed spectrum, refs 12 and 13 found that the spectrum appears surprisingly blue, even if no additional Lyman continuum absorption is assumed to take place in the galaxy (beyond the dust absorption that also affects the spectrum at 1,500 Å).

We compare the stacked composite spectrum of LBGs in ref. 1 with

model galactic spectra, including metal-free stars. The intrinsic emission of the galaxies is modelled as the sum of ‘normal’ star formation (stellar populations with metallicity ratio $>10^{-3}Z/Z_{\odot}$; refs 14, 15), and metal-free, massive population III star formation. (Subscript $\odot$ refers to the Sun.) For the ‘normal’ component of the intrinsic stellar emission, we use the spectral synthesis models of ref. 16 (see <http://www.stsci.edu/science/starburst99/>) for electronic data) for a Salpeter IMF (with a slope $\alpha = 2.35$) between 1–100 $M_{\odot}$ and a metallicity of $Z = 0.4Z_{\odot}$, and compute the spectrum for continuous star formation for 10⁸ yr, which is the age at which a steady state has been reached. For the ‘metal-free’ component, we use the models of ref. 17, and adopt the same Salpeter IMF, but limited to massive stars in the 100–500 $M_{\odot}$ range^{18,19}, and with $Z = 0$. Note that massive stars with $>100M_{\odot}$ have similar spectra, and our results should not depend significantly on the distribution of population III stellar masses above this threshold value. To simulate the observed, processed composite spectrum, we modify the emitted stellar template spectrum by including the effects of dust absorption, as well as the opacity of the intervening intergalactic medium (see Supplementary Information).

As the top panel in Fig. 1 shows, in the $880 \text{ Å} \leq \lambda \leq 910 \text{ Å}$ range, the mean observed flux is quite high, $F_{\nu}/F_{1,500} = 0.064 \pm 0.013$; while the normal-only star-formation model predicts a much lower value of $F_{\nu}/F_{1,500} = 0.023 \pm 0.01$ (here F_{ν} and $F_{1,500}$ represent flux at frequency ν and wavelength 1,500 Å, respectively). Taken at face value, the data reveals an ionizing flux about three times higher than this model predicts, even if we do not include any continuum absorption by neutral gas in the galaxy. The data and the model do not agree at the $>3\sigma$ level. The model with population III star formation with $f_{\text{popIII}} = 0.5$ almost generates the observed ionizing flux in the $880 \text{ Å} \leq \lambda \leq 910 \text{ Å}$ range; producing a mean $F_{\nu}/F_{1,500} = 0.048 \pm 0.02$.

Our conclusions regarding the LBGs apply to the bluest subsample. This is further supported by the analysis of the C IV absorption line associated with the stellar winds in massive stars that is observed in the spectra of LBGs. According to our prediction, the bluest LBGs should have weaker C IV absorption than the reddest LBGs, because population III stars are metal-free and therefore contribute significantly fewer metals. This trend is indeed what is observed (see table 3 and section 5.3 in ref. 3). Detailed inspection of table 3 of ref. 3 shows that the weaker strength of the C IV line requires an abundance of $\sim 50\%$ population III stars in the bluest LBGs. We note that the C IV profile in LBGs is a complex superposition of interstellar absorption, stellar-wind emission and absorption, and possibly nebular emission. Our models predict that the trend of decreasing C IV strength in the bluest galaxies is present in the narrow interstellar absorption and any broad stellar component (which measures the total cumulative C output of all stars), but not necessarily in any of the other components. The fact that metal

¹Department of Physics and Astronomy, University of Pennsylvania, 209 South 33rd Street, Philadelphia, Pennsylvania 19104, USA. ²Department of Astronomy, Columbia University, 550 West 120th Street, New York, New York 10027, USA.

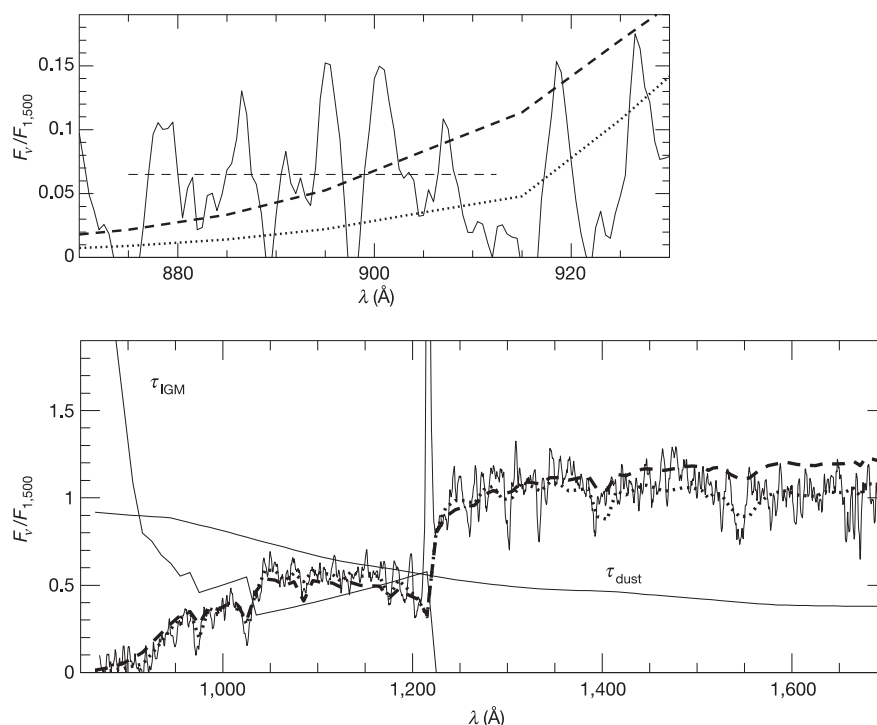


Figure 1 | Lyman break galaxies stacked observed spectrum and best-fit population synthesis models. The dotted curve corresponds to continuous normal star formation for 10^8 years, and the dashed curve corresponds to a mixed model in which 50% of the stars (by mass) are massive, metal-free stars in the $100\text{--}500M_{\odot}$ range. The solid curves show the composite spectrum in ref. 1. All fluxes are normalized to the flux at the emitted wavelength of $1,500\text{ Å}$. For reference, in the bottom panel, we show the optical depths τ assumed for the dust and the IGM. In the top panel, the horizontal dashed line shows the mean observed flux in the interval $880\text{ Å} \leq \lambda \leq 910\text{ Å}$. The inference is that the spectrum requires a significant contribution from metal-free star formation, with a mass fraction of $f_{\text{popIII}} \approx 0.5$. As discussed in ref. 12, changing the assumed metallicity of the

normal stars does not significantly increase the predicted ionizing fluxes. The ultraviolet fluxes of OB stars can be increased by the presence of stellar winds (not included in the stellar models we adopted). Although winds can increase the He II-ionizing flux by orders of magnitude, the corresponding increase for the H-ionizing flux in O stars has been found to be small²⁷. The increase can be more significant for cooler B stars²⁸, but these stars do not dominate the ionizing photon budget in the continuous star-formation models, in which O stars are continuously replenished. Furthermore, the hydrostatic stellar models typically reproduce the properties of galactic H II regions¹⁶, so that an increase by the required factor of about two would make it more difficult to reconcile the models with these observations.

lines are observed in all LBGs indicates that LBGs are not composed purely of population III stars, but that there is a mixture in each galaxy of population III and normal stars.

There have been recent observations of a population of 35 Ly α blobs at $z = 3.1$ in narrow and broadband spectra². For about one-third of the blobs, the number of photons provided by a normal stellar population that fits the observed spectrum at wavelengths longer than 912 Å is not capable of producing the required number of ionizing photons to explain the observed Ly α flux. If we assume case B (two Ly α photons are produced for every three ionizing photons from the stars) then a stellar population with $Z = 10^{-2}Z_{\odot}$ would produce at least 35% fewer photons than those necessary to account for the observed Ly α flux. To reproduce the observed flux, about 30% of the stars would need to be metal-free.

Figure 2 shows that about 20–30% of the blobs also have continuum colours consistent with the stars that produce the ionizing Ly α photons being metal-free. The fact that the stellar populations powering the extended emission are metal-poor is consistent with the fact that these sources are extended, and are presumably more likely to have been caught in the process of their assembly, when they are less metal-enriched overall. They could plausibly be identified, therefore, as the progenitors of LBGs. The rest-frame equivalent width for Ly α ranges from 20 to 350 Å . For models with constant star formation and $Z > 10^{-5}Z_{\odot}$ the equivalent widths are $70\text{--}100\text{ Å}$. About one-third of the blobs have equivalent widths larger than 100 Å and therefore require metal-free stars. This is in agreement with what is found for the colours. We note that the good fit to the $B - I$ colours, without including (case B) reprocessing of ionizing

radiation into Lyman α , implies that f_{esc} is of the order of 100%, because a smaller escape fraction would imply more flux on B and therefore significantly bluer $B - I$ colours than those observed. The large escape fraction of ionizing radiation would be consistent with the value observed in the composite of LBG spectra.

Strong He II ($1,640\text{ Å}$) emission line from a composite of LBGs has been observed³. They report equivalent widths for a sample of several hundred LBGs of about 2 Å . This line is also very broad, with a full-width at half-maximum of about $1,500\text{ km s}^{-1}$. Although this cannot be explained by stars without winds, it has been shown (see refs 20, 21) that, aided by rotation, metal-free stars can lose significant amounts of mass and drive strong winds, similar to normal-metallicity stars. In this case, the He-ionizing photons from the massive population III stars will be reprocessed in the optically thick stellar outflows surrounding the star, making the line broad²². As discussed in ref. 4, normal Wolf-Rayet stars cannot explain the He II line observed in the composite LBG sample, because the requisite number of Wolf-Rayet stars will overproduce the stellar C IV emission.

However, metal-free massive stars with winds do not have this problem. As shown in ref. 22, there is a clear anti-correlation between the strength of the He II $1,640\text{ Å}$ line and the C IV line (see figure 13 in ref. 22 for the case of $Z = 0.01$ and $10^{-4}Z_{\odot}$). For metal-free stars the equivalent width of the He II $1,640\text{ Å}$ line for a continuous star-forming ($1M_{\odot}\text{ yr}^{-1}$) model of metal-free stars at age 0.1 Gyr is 30 Å , significantly larger than observed. So for this sample, the required amount of metal-free stars is only 10%. However, the equivalent width of the He II $1,640\text{ Å}$ line is very sensitive to the metallicity. If we use

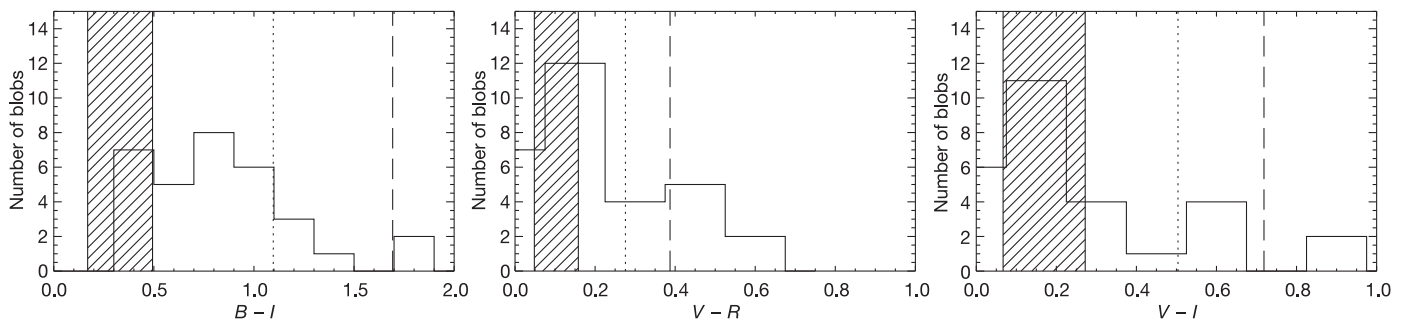


Figure 2 | Colour distributions of the Ly α blobs. The different lines correspond to stellar population models with continuous star formation at an age of 10^8 yr for $Z = 0.2Z_{\odot}$ and $Z = 0.01Z_{\odot}$ (dashed and dotted lines respectively; both computed using the models from ref. 29) and $Z = 0$ (solid line, assuming an IMF and mass range $100\text{--}500M_{\odot}$ as in our analysis of LBG spectra). We have included the effect of the opacity of the intervening

intergalactic medium. The shaded ranges indicate the range of expected colours as various properties of the metal-free population are varied, such as their age (between $10^5\text{--}10^8$ yr), the IMF (for the three models given in ref. 17), and star-formation history (between a burst and continuous star formation at a constant rate). All three colour histograms suggest that about 20% of the blobs need to be made of purely population III stars.

$Z = 10^{-7}Z_{\odot}$ (instead of $Z = 0$), then the $912\text{ \AA}$ decrement does not change, so our conclusions above about the ultraviolet radiation from LBGs and the Ly α blobs remain unchanged. But the He-ionizing flux, and the He II $1,640$ equivalent width is reduced by about a factor of 6–7. This implies a population of $\sim 30\%$ metal-free stars.

Recent non-detections of the He II $1,640$ line in a composite spectrum of 17 galaxies at $z = 4.5$ (ref. 23), as well as from an individual Ly α emitter at $z = 6.33$ (ref. 24) have provided tight upper limits, but still allow for a significant mass fraction ($\sim 50\%$) in population III stars.

The Large Area Lyman Alpha survey³ reports large Ly α equivalent widths for a sample of 150 emitters at $z \approx 4.5$. The median value of the distribution is $\sim 450\text{ \AA}$ and 60% have equivalent widths above $240\text{ \AA}$. No continuum emission was found in 30% of the emitters. However, the equivalent widths are spread over a wide range: from 10 to $5,000\text{ \AA}$. This implies that no single type of stellar population can account for the large spread of equivalent widths. The first suggestion that these large equivalent widths can be produced by very-low-metallicity stars is found in ref. 30. We find that for metal-free stars and constant star formation, the equivalent width is $1,000\text{ \AA}$ for the IMF used in this paper and declines to $300\text{ \AA}$ for a Salpeter IMF (ref. 17). We performed a more detailed fit to the distribution in ref. 3. Using models for metallicities $Z > 10^{-5}Z_{\odot}$ we find that bursts of ages greater than 10^7 years have equivalent widths as large as $300\text{ \AA}$, whereas constant star formation models have about $200\text{ \AA}$. Therefore ‘normal’ stellar populations can account for as much as 40% of the Ly α emitters, while the rest need to be accounted for by metal-free stellar populations.

Recently, two stars with metallicity $Z < 10^{-5}Z_{\odot}$ have been found in the Milky Way²⁵. Although such low-mass stars ($\leq 0.8M_{\odot}$) are not included in our model, they reveal the details of the earliest stages of chemical enrichment, and suggest an initial overproduction of carbon. For stars with $[\text{Fe}/\text{H}] < -3.5$, 40% have $[\text{C}/\text{Fe}] > 1.0$ (ref. 26) and the trend increases to 100% for stars with $[\text{Fe}/\text{H}] < -5$. This trend can naturally arise in our model^{20,21}.

Our results imply that micro-mixing of metals within galaxies is inefficient on a gigayear timescale, a conclusion that may be surprising, as current models of hierarchical galaxy formation assume efficient intragalactic metal mixing, and therefore do not predict metal-free star formation significantly below $z \approx 5$ (refs 5–9). However, the presence of significant pockets of metal-free star formation explains four puzzling observations, and this hypothesis does not violate any other existing observations. The presence of metal-free gas pockets can be verified in future high-resolution numerical simulations, and by observation. Our models predict that in some of the LBGs there should be a detectable continuum shortward of $912\text{ \AA}$ and a detectable individual He II line. Long-integration observations of these individual galaxies in the future

could be used to characterize their spectra, and to confirm the presence of metal-free stellar populations. In particular, we predict that there should be a clear anti-correlation between He II $1,640$ and C IV line strength and this should be a reliable test of the amount of metal-free stars at $z \approx 3\text{--}5$.

Received 6 September 2004; accepted 4 January 2005.

- Steidel, C. C., Pettini, M. & Adelberger, K. L. Lyman-continuum emission from galaxies at $z \sim 3.4$. *Astrophys. J.* **546**, 665–671 (2001).
- Matsuda, Y. et al. A Subaru search for Ly α blobs in and around the protocluster region at redshift $z = 3.1$. *Astrophys. J.* **128**, 569–584 (2004).
- Malhotra, S. & Rhoads, J. E. Large equivalent width Ly α line emission at $z = 4.5$: young galaxies in a young Universe? *Astrophys. J.* **565**, L71–L74 (2002).
- Shapley, A. E., Steidel, C. C., Pettini, M. & Adelberger, K. L. Rest-frame ultraviolet spectra of $z \sim 3$ Lyman break galaxies. *Astrophys. J.* **588**, 65–89 (2003).
- Madau, P., Ferrara, A. & Rees, M. J. Early metal enrichment of the intergalactic medium by pregalactic outflows. *Astrophys. J.* **555**, 92–105 (2001).
- Scannapieco, E., Schneider, R. & Ferrara, A. The detectability of the first stars and their cluster enrichment signatures. *Astrophys. J.* **589**, 35–52 (2003).
- Norman, M. L., O’Shea, B. W. & Paschos, P. Did massive primordial stars pre-enrich the Ly α forest? *Astrophys. J.* **601**, L115–L118 (2004).
- Yoshida, N., Bromm, V. & Hernquist, L. The era of massive Population III stars: cosmological implications and self-termination. *Astrophys. J.* **605**, 579–590 (2004).
- Bromm, V. & Loeb, A. High-redshift gamma-ray bursts from Population III progenitors. Preprint at (<http://arXiv.org/astro-ph/0509303>) (2005).
- Tumlinson, J. & Shull, J. M. Zero-metallicity stars and the effects of the first stars on reionization. *Astrophys. J.* **528**, L65–L69 (2000).
- Bromm, V., Kudritzki, R. P. & Loeb, A. Generic spectrum and ionization efficiency of a heavy initial mass function for the first stars. *Astrophys. J.* **552**, 464–472 (2001).
- Hui, L., Haiman, Z., Zaldarriaga, M. & Alexander, T. Connections between the cosmic baryon fraction, the extragalactic ionizing background, and Lyman break galaxies. *Astrophys. J.* **564**, 525–533 (2002).
- Haehnelt, M., Madau, P., Kudritzki, R. & Haardt, F. An ionizing ultraviolet background dominated by massive stars. *Astrophys. J.* **549**, L151–L154 (2001).
- Bromm, V., Ferrara, A., Coppi, P. S. & Larson, R. B. The fragmentation of pre-enriched primordial objects. *Mon. Not. R. Astron. Soc.* **328**, 969–976 (2001).
- Schneider, R., Ferrara, A., Natarajan, P. & Omukai, K. First stars, very massive black holes, and metals. *Astrophys. J.* **571**, 30–39 (2002).
- Leitherer, C. et al. Starburst 99: Synthesis models for galaxies with active star formation. *Astrophys. J. Suppl.* **123**, 3–30 (1999).
- Schaerer, D. The transition from Population III to normal galaxies: Ly α and He II emission and the ionizing properties of high redshift starburst galaxies. *Astron. Astrophys.* **397**, 527–538 (2003).
- Abel, T., Bryan, G. L. & Norman, M. L. The formation of the first star in the Universe. *Science* **295**, 93–95 (2002).
- Bromm, V., Coppi, P. S. & Larson, R. B. The formation of the first stars. I. The primordial star-forming cloud. *Astrophys. J.* **564**, 23–51 (2002).
- Marigo, P. et al. Zero-metallicity stars. II. Evolution of very massive objects with mass loss. *Astron. Astrophys.* **399**, 617–630 (2003).
- Meynet, G., Ekström, S. & Maeder, A. The early star generations: the dominant effect of rotation on the CNO yields. Preprint at (<http://arXiv.org/astro-ph/0510560>) (2005).

22. Kudritzki, R. P. Line-driven winds, ionizing fluxes, and ultraviolet spectra of hot stars at extremely low metallicity. I. Very massive O stars. *Astrophys. J.* **577**, 389–408 (2002).
23. Dawson, S. *et al.* Spectroscopic properties of the $z = 4.5$ Ly α emitters. *Astrophys. J.* **617**, 707–717 (2004).
24. Nagao, T. *et al.* An observational pursuit for population III stars in a Ly α emitter at $z = 6.33$ through He II emission. *Astrophys. J.* **631**, L5–L8 (2005).
25. Christlieb, N. *et al.* A stellar relic from the early Milky Way. *Nature* **419**, 904–906 (2002).
26. Beers, T. C. & Christlieb, N. The discovery and analysis of very metal-poor stars in the galaxy. *Annu. Rev. Astron. Astrophys.* **43**, 531–580 (2005).
27. Schaerer, D. & de Koter, A. Combined stellar structure and atmosphere models for massive stars. III. Spectral evolution and revised ionizing fluxes of O3–B0 stars. *Astron. Astrophys.* **322**, 598–610 (1997).
28. Najarro, F., Kudritzki, R. P., Cassinelli, J. P., Stahl, O. & Hillier, D. J. Stellar winds and the EUV continuum excess of early B-giants. *Astron. Astrophys.* **306**, 890–892 (1996).
29. Jimenez, R., MacDonald, J., Dunlop, J. S., Padoan, P. & Peacock, J. A. Synthetic stellar populations: single stellar populations, stellar interior models and primordial protogalaxies. *Mon. Not. R. Astron. Soc.* **349**, 240–254 (2004).
30. Kudritzki, R. P. *et al.* Discovery of nine Ly α emitters at redshift $z \sim 3.1$ using narrowband imaging and VLT spectroscopy. *Astrophys. J.* **536**, 19–30 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Y. Matsuda and T. Yamada for providing the unpublished raw colour data that appears in Fig. 2, D. Schaerer for providing pop-III spectra in electronic form, and A. Shapley for useful discussions. We also thank T. Beers for pointing out the C trends in low metallicity stars and suggesting that these could be explained by our model. R.J. and Z.H. gratefully acknowledge financial support from NSF and NASA.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.J. (raulj@physics.upenn.edu) or Z.H. (zoltan@astro.columbia.edu).

A non-spherical core in the explosion of supernova SN 2004dj

Douglas C. Leonard^{1†}, Alexei V. Filippenko², Mohan Ganeshalingam², Franklin J. D. Serduke², Weidong Li², Brandon J. Swift², Avishay Gal-Yam¹, Ryan J. Foley², Derek B. Fox^{1†}, Sung Park², Jennifer L. Hoffman² & Diane S. Wong²

An important and perhaps critical clue to the mechanism driving the explosion of massive stars as supernovae is provided by the accumulating evidence for asymmetry in the explosion. Indirect evidence comes from high pulsar velocities¹, associations of supernovae with long-soft γ -ray bursts^{2,3}, and asymmetries in late-time emission-line profiles⁴. Spectropolarimetry provides a direct probe of young supernova geometry, with higher polarization generally indicating a greater departure from spherical symmetry^{5,6}. Large polarizations have been measured for ‘stripped-envelope’ (that is, type Ic; ref. 7) supernovae, which confirms their non-spherical morphology^{8,9}; but the explosions of massive stars with intact hydrogen envelopes^{7,10} (type II-P supernovae) have shown only weak polarizations at the early times observed^{11,12}. Here we report multi-epoch spectropolarimetry of a classic type II-P supernova that reveals the abrupt appearance of significant polarization when the inner core is first exposed in the thinning ejecta (~ 90 days after explosion). We infer a departure from spherical symmetry of at least 30 per cent for the inner ejecta. Combined with earlier results, this suggests that a strongly non-spherical explosion may be a generic feature of core-collapse supernovae of all types, where the asphericity in type II-P supernovae is cloaked at early times by the massive, opaque, hydrogen envelope.

SN 2004dj was discovered¹³ on 31 July 2004 (UT) in the nearby spiral galaxy NGC 2403, and was quickly identified¹⁴ as a core-collapse event with spectral features similar to those of the type II-P supernova SN 1999em (ref. 10) about three weeks after explosion. At an estimated distance of 3.13 Mpc (ref. 15), it is the closest normal type II-P supernova observed to date. We present spectropolarimetry spanning the first nine months of its development in Table 1 and in Figs 1 and 2.

The light curve of SN 2004dj (Fig. 2) exhibits the classic features of a type II-P supernova: an enduring period of nearly constant optical luminosity followed by a precipitous drop in brightness and a smooth decline on the radioactive tail. The first three epochs of spectropolarimetry, obtained during the middle and end of the plateau phase, find the supernova to be unpolarized (Fig. 1), indicating a substantially spherical geometry for the electron-scattering atmosphere while the photosphere is receding back through the hydrogen envelope.

A global asphericity of the electron-scattering atmosphere will manifest itself through two main spectropolarimetric signatures. First, there is significant polarization in spectral regions not dominated by line opacity, resulting from a non-balancing of the electric vectors of the scattered continuum photons^{5,8}. For type II-P supernovae, this includes the broad spectral region 6,800–8,200 Å (ref. 16).

Very little polarization should be seen in spectral regions shortward of wavelength $\lambda \approx 5,500$ Å owing to line-blanking by Fe; other spectral regions with strong individual emission lines (for example, H α and Ca II near-infrared triplet) should also be relatively unpolarized¹⁷. Second, polarization increases are anticipated in the troughs of strong P-Cygni absorption lines^{11,18}. This results from selective blocking of more forward-scattered and, hence, less polarized, light in the trough regions. A further expectation resulting from a global asphericity, as opposed to other potential supernova polarization mechanisms (see ref. 19 for an extensive discussion), is that no rotation of the polarization angle (PA) should be seen across these absorption-line features.

Once the ionized hydrogen envelope has recombined, the optical brightness of a type II-P supernova drops dramatically as the

Table 1 | Spectropolarimetric observations of SN 2004dj

Epoch (d)	Total exposure (s)	p (%) $\pm 1\sigma$	θ (deg) $\pm 1\sigma$
39	600	0.079 ± 151	ND
58	9,200	0.000 ± 041	ND
72	2,360	-0.006 ± 041	ND
91	5,200	0.558 ± 042	28 ± 2
128	3,600	0.401 ± 054	178 ± 4
157	19,200	0.215 ± 042	173 ± 6
187	14,400	0.141 ± 052	178 ± 11
242	9,200	0.071 ± 047	171 ± 20
271	5,600	0.052 ± 058	ND

Epochs are with respect to the estimated date of explosion, 2004 July 14 (ref. 14). Continuum polarization level, p , and polarization position angle in the plane of the sky, θ , represent measurements made on the ISP-subtracted data (see Fig. 1 legend) over the range 6,800–8,200 Å; in an effort to avoid statistical biases that exist at low polarization levels, the reported polarization is the median of the rotated Stokes parameter¹¹ over this interval. The PA is determined as $\theta = 1/2 \times \tan^{-1}(u_{\text{med}}/q_{\text{med}})$, where u_{med} and q_{med} are the median values of the Stokes parameters over this interval. The day 39 data were acquired using the double spectrograph with polarimeter on the 5-m Hale Telescope at Palomar Observatory. All other data were obtained with the Kast double spectrograph with polarimeter at the Cassegrain focus of the Shane 3-m telescope at Lick Observatory. The uncertainties in the Stokes parameters, which were then used to derive the uncertainties in p and θ , are derived as the quadrature sum of the Poisson error and a term reflecting our estimate of the random uncertainty in the overall, measured level of the Stokes parameters. For the Lick data, we set this term to be 0.04%, which is the average 1σ (s.d.) spread in the measured Stokes parameters of null standards observed on multiple nights (the same nights as the SN 2004dj data were obtained). For the day 39 data obtained at Palomar, we set this term equal to 0.15%, to account for the likely existence of some instrumental polarization in the data: while null standards at Lick were always found to be null to within 0.1%, the null standards observed at Palomar on this night possessed measured polarizations of up to 0.25%, in a way that could not be consistently modelled or removed from the data. Note that values for θ are ill-defined for the day 39, 58, 72 and 271 data owing to the low polarization. On day 271, a final 2,367 seconds of observations (a third complete set of four waveplate positions) were acquired but not included when forming the combined Stokes parameters for the night's observations, as they were taken under deteriorating weather conditions and with poor observing technique (object accidentally placed near the edge of the slit), and are discrepant with the earlier data at the 4σ (s.d.) level. ND, not determined.

¹Astronomy Department, MS 105-24, California Institute of Technology, Pasadena, California 91125, USA. ²Department of Astronomy, University of California, Berkeley, California 94720-3411, USA. [†]Present addresses: Department of Astronomy, San Diego State University, San Diego, California 92182, USA (D.C.L.); Department of Astronomy and Astrophysics, Pennsylvania State University, 525 Davey Lab, University Park, Pennsylvania 16802, USA (D.B.F.).

recombination front rapidly recedes through the helium core¹⁶. Previous studies¹² have noted the confluence of factors making the drop-off of the plateau a propitious period in which to search for evidence of explosion asphericity, as the innermost regions are exposed at a point when sufficient optical depth to electron scattering in the inner core still exists to polarize the light and, hence, reveal the geometry. Our fourth spectropolarimetric observation captures this moment in a type II-P supernova (to our knowledge, for the first time). Taken during the steepest part of the descent off the plateau (Fig. 2), on day 91 after explosion, the data show a dramatic change from those obtained just 19 days earlier: the polarization level has increased to nearly 0.6% in the continuum-dominated region 6,800–8,200 Å (Fig. 1), as well as in the troughs of some of the stronger P-Cygni line features, including most prominently that due to Na I D. The PAs in the P-Cygni troughs agree with the continuum PA, with a measured value of $\theta = 32^\circ \pm 3^\circ$ in the Na I D absorption compared with $\theta = 28^\circ \pm 2^\circ$ for the continuum region. The observed spectropolarimetric behaviour of SN 2004dj matches expectations resulting from a strong departure from spherical symmetry for the electron-scattering atmosphere in the inner regions of

the ejecta (expansion velocities less than $\sim 2,500 \text{ km s}^{-1}$, or about 1/10 the radius of the supernova assuming a maximum ejecta velocity of $\sim 25,000 \text{ km s}^{-1}$).

As SN 2004dj moves into the nebular phase and the atmosphere continues to expand, the optical depth to electron scattering, and thus the polarization, is expected to fall. Specifically, if the day 91 observation occurred at or beyond the time when the single-scattering limit is reached in the supernova's atmosphere (a single-scattering atmosphere produces maximum supernova polarization⁶), then the polarization should subsequently decrease in proportion to the (angle-averaged) electron-scattering optical depth, as this determines the percentage of scattered light. For this simple scenario, the polarization would be expected to decline with time t as $1/t^2$, owing to geometrical dilution caused by homologous expansion (that is, the same number of scattering electrons spread over a cross-sectional area that grows as t^2). This prediction is borne out rather well by the data (Fig. 2), although there is some evidence suggestive of a more complex situation. Specifically, the PA changes by 30° between days 91 and 128, after which it remains constant for the next three epochs (Table 1). This result has a high degree of statistical significance, and indicates that the scattering environment's morphology might be more complicated than that of a single, fixed geometry; we speculate on possible causes of this PA rotation below.

The link between long-soft γ -ray bursts (GRBs) and a subset of type Ic supernovae—specifically those with spectral lines indicating very high expansion velocities at early times⁴—is now secure^{2,3}, and successful models for GRBs require a jet-like explosion from energetics considerations. While it is not clear that the same mechanism that generates the GRB is also responsible for exploding the star, the substantial early-time polarization that has been measured for some of these events⁹, and for more typical type Ic supernovae as well⁸, supports the contention of a non-spherical explosion for 'stripped-envelope' supernovae. The most thoroughly studied non-spherical

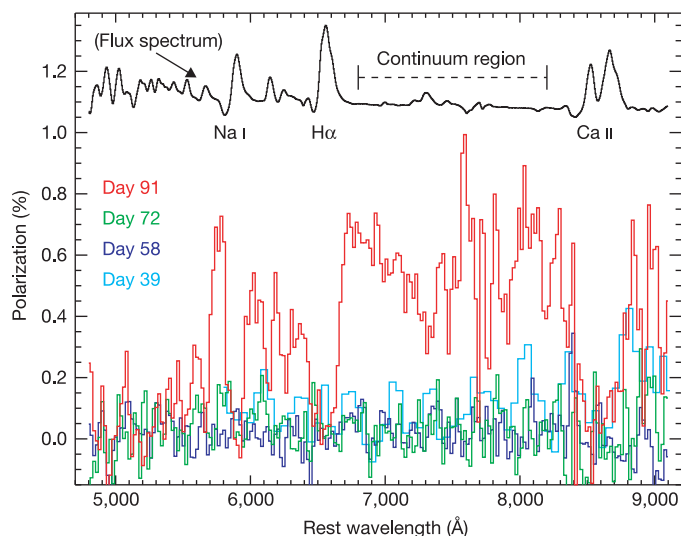


Figure 1 | Intrinsic polarization of SN 2004dj during the photospheric phase, with day since explosion indicated. The flux spectrum is from day 91, with prominent line features labelled. The recession velocity of 131 km s^{-1} (NASA Extragalactic Database) has been removed from all data. The polarization data have been corrected for an interstellar polarization (ISP) contribution, which was derived by examining (1) the observed polarization in spectral regions dominated by the strongest emission lines (that is, Na I D, H α and the Ca II near-infrared triplet), which are believed to be intrinsically unpolarized¹⁷, and (2) the observed polarization at late times (that is, the nebular phase), when the electron-scattering optical depth of the atmosphere has dropped well below unity and the observed polarization at all wavelengths is largely due to ISP (see ref. 19 for a complete discussion). These complementary approaches yield virtually identical results (to within 0.1%) for the ISP estimation. For our data set, the final epoch is from day 271, at which point the ejecta are probably not fully nebular and some small intrinsic polarization ($<0.1\%$) may still exist. Examining the temporal evolution in the Stokes q - u plane, we find the late-time polarization to be steadily decreasing down to the level indicated by the earliest measurements made during the plateau phase, suggesting that these early epochs possess little to no intrinsic polarization (this is also supported by the lack of any line features at these times). Thus, to derive the intrinsic polarization at all epochs, we fitted a low-order spline to the Stokes q and u spectra of the data from day 58 (note that the data from day 39 have significantly lower signal-to-noise ratio than those obtained at the other plateau epochs), and then subtracted the result from the Stokes parameters obtained on all other epochs. This resulted in the removal of $p_{\text{ISP}} \approx 0.29\%$, $\theta_{\text{ISP}} \approx 20^\circ$ from the observed data.

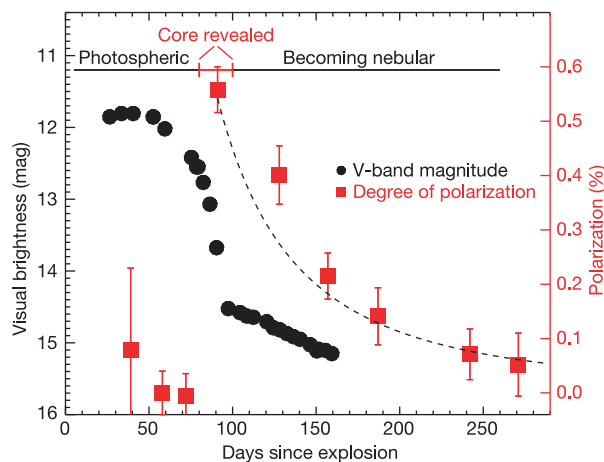


Figure 2 | Light curve and continuum polarization of SN 2004dj.

Polarization measures are from Table 1; error bars are 1σ (s.d.) statistical. Photometry is from data obtained with the 30-inch (0.8-m) Katzman Automatic Imaging Telescope (KAIT) at Lick Observatory and the 60-inch (1.5-m) telescope at Palomar Observatory. The dashed line represents the expected decline in polarization ($p = (t_0/t)^2 \times p_0$, where $t_0 = 91$ days and $p_0 = 0.558\%$ from Table 1) during the transition to the nebular phase due to the effects of diminishing electron scattering in optically thin, expanding ejecta. Note that the age of SN 2004dj at discovery and, hence, the plateau duration, is not well constrained by direct observation, as NGC 2403 had just emerged from solar conjunction and the most recent reported pre-explosion image was taken over six months earlier²⁷. Our adopted explosion date of 2004 July 14 results from the spectral analysis of ref. 14, but we note that the light-curve modelling of ref. 24 yields an explosion date 31 days earlier. If the earlier date were adopted, the estimated plateau duration would increase from ~ 70 to ~ 100 days.

explosion model is that of bipolar jets²⁰, which distribute radioactive ⁵⁶Ni and other synthesized heavy elements (such as Fe) preferentially along the polar axes; lighter elements produced by the progenitor star during its evolution (for example, O) are distributed near the equatorial plane in a disk-like structure²¹. This model, in fact, has recently been successfully used to explain the emergence of double-peaked oxygen line profiles in spectra of SN 2003jd, a type Ic supernova with high early-time expansion velocities, obtained a year after explosion⁴.

For type II-P supernovae, the bipolar jet model predicts the radioactive ⁵⁶Ni to be confined to the inner regions, where it would ionize the surrounding material, produce a non-spherical electron-scattering photosphere when exposed²², and plausibly produce polarization similar to that observed in SN 2004dj. In addition, it has been argued²³ that a non-spherical distribution of ⁵⁶Ni can produce asymmetries in the H α emission-line profile of type II-P supernovae during the nebular phase. Such asymmetries have, in fact, been recently identified in late-time spectra of SN 2004dj, and are interpreted within the context of the bipolar jet model by ref. 24, which concludes that they are best explained by a bipolar distribution of ⁵⁶Ni with the major axis oriented $30^\circ \pm 10^\circ$ to our line-of-sight. If we adopt this viewing orientation, then our measured polarization peak of 0.56% places a minimum departure from spherical symmetry of $\sim 30\%$ (axis ratio 10:7) for the inner regions of SN 2004dj from the axially symmetric, electron-scattering models of ref. 6. Naturally, if the electron-scattering atmosphere on day 91 has not yet reached the single-scattering limit, or is already below it, then the actual degree of asphericity is greater than our derived value.

The lack of polarization during the plateau fits comfortably within the bipolar jet model as well, as hydrodynamical simulations demonstrate that the shock propagates laterally as it penetrates the inner regions²¹, and ultimately becomes completely spherical as it proceeds through the massive hydrogen envelope²⁵. Because the shock propagation (and, hence, ⁵⁶Ni production) in the bipolar jet model is not purely radial in the inner regions, a rapidly but smoothly changing PA might be expected during the brief period (SN 2004dj falling off the optical plateau) in which the ⁵⁶Ni is being uncovered. This could explain the PA rotation observed for SN 2004dj between days 91 and 128, although opacity effects may also play a role²⁶. Additional examples, preferably with fine temporal sampling during the critical phase at the end of the plateau, are needed to fully test these conjectures.

Analysis of pre-explosion images places the initial main-sequence mass of SN 2004dj's progenitor at 12 solar masses²⁷, in accord with masses inferred by similar studies of other type II-P supernovae (ref. 28 and references therein). Theory predicts that single stars that explode as type Ic supernovae must have initial main-sequence masses greater than ~ 25 solar masses to undergo the extraordinary wind-driven mass loss required to remove their entire H/He envelopes before explosion²⁹, and observations of Galactic Wolf-Rayet stars support this conclusion³⁰. The prompt appearance of substantial polarization in type Ic supernovae supports a non-spherical explosion model for these very massive stars. The discovery of a non-spherical core in SN 2004dj suggests that a similar explosion mechanism is at work in type II-P supernovae, and may, in fact, be inherent to the core-collapse process, regardless of progenitor mass.

Received 19 October; accepted 15 December 2005.

1. Chatterjee, S. *et al.* Getting its kicks: A VLBA parallax for the hyperfast pulsar B1508+55. *Astrophys. J.* **630**, L61–L64 (2005).
2. Galama, T. J. *et al.* An unusual supernova in the error box of the gamma-ray burst of 25 April 1998. *Nature* **395**, 670–672 (1998).

3. Matheson, T. *et al.* Photometry and spectroscopy of GRB 030329 and its associated supernova 2003dh: The first two months. *Astrophys. J.* **599**, 394–407 (2003).
4. Mazzali, P. A. *et al.* An asymmetric, energetic type Ic supernova viewed off-axis and a link to gamma-ray bursts. *Science* **308**, 1284–1287 (2005).
5. Shapiro, P. R. & Sutherland, P. G. The polarization of supernova light—A measure of deviation from spherical symmetry. *Astrophys. J.* **263**, 902–924 (1982).
6. Höflich, P. Asphericity effects in scattering dominated photospheres. *Astron. Astrophys.* **246**, 481–489 (1991).
7. Filippenko, A. V. Optical spectra of supernovae. *Annu. Rev. Astron. Astrophys.* **35**, 309–355 (1997).
8. Leonard, D. C. & Filippenko, A. V. in *Supernovae as Cosmological Lighthouses* (eds Turatto, M., Benetti, S., Zampieri, L. & Shea, W.) 330–336 (ASP, San Francisco, 2005).
9. Wang, L., Baade, D., Höflich, P. A. & Wheeler, J. C. Spectropolarimetry of the Type Ic supernova SN 2002ap in M74: More evidence for asymmetric core collapse. *Astrophys. J.* **592**, 457–466 (2003).
10. Leonard, D. C. *et al.* The distance to SN 1999em in NGC 1637 from the expanding photosphere method. *Publ. Astron. Soc. Pacif.* **114**, 35–64 (2002).
11. Leonard, D. C., Filippenko, A. V., Ardila, D. R. & Brotherton, M. S. Is it round? Spectropolarimetry of the Type II-P supernova 1999em. *Astrophys. J.* **553**, 861–885 (2001).
12. Leonard, D. C. & Filippenko, A. V. Spectropolarimetry of the Type II supernovae 1997ds, 1998A, and 1999gi. *Publ. Astron. Soc. Pacif.* **113**, 920–936 (2001).
13. Nakano, S., Itagaki, K., Bouma, R. J., Lehy, M. & Hornoch, K. Supernova 2004dj in NGC 2403. *IAU Circ.* **8377** (2004).
14. Patat, F., Benetti, S., Pastorello, A., Filippenko, A. V. & Aceituno, J. Supernova 2004dj in NGC 2403. *IAU Circ.* **8378** (2004).
15. Freedman, W. L. *et al.* Final results from the Hubble Space Telescope key project to measure the Hubble constant. *Astrophys. J.* **553**, 47–72 (2001).
16. Eastman, R. G., Woosley, S. E., Weaver, T. A. & Pinto, P. A. Theoretical light curve of a Type II-P supernova. *Astrophys. J.* **430**, 300–310 (1994).
17. Höflich, P., Wheeler, J. C., Hines, D. C. & Trammell, S. R. Analysis of the polarization and flux spectra of SN 1993. *J. Astrophys. J.* **459**, 307–321 (1996).
18. Jeffery, D. J. Analysis of SN 1987A polarimetry. *Astrophys. J.* **375**, 264–287 (1991).
19. Leonard, D. C., Li, W., Filippenko, A. V., Foley, R. J. & Chornock, R. Evidence for spectropolarimetric diversity in Type Ia supernovae. *Astrophys. J.* **632**, 450–475 (2005).
20. Wheeler, J. C., Meier, D. L. & Wilson, J. R. Asymmetric supernovae from magnetocentrifugal jets. *Astrophys. J.* **568**, 807–819 (2002).
21. Maeda, K. *et al.* Explosive nucleosynthesis in aspherical hypernova explosions and late-time spectra of SN 1998bw. *Astrophys. J.* **565**, 405–412 (2002).
22. Höflich, P. A., Baade, D., Khokhlov, A. M., Wang, L. & Wheeler, J. C. in *A Massive Star Odyssey: from Main Sequence to Supernova* (eds van der Hucht, K. A., Herrero, A. & Esteban, C.) Proc. IAU Symp. 212 387–394 (ASP, San Francisco, 2003).
23. Elmhamdi, A. *et al.* Photometry and spectroscopy of the Type IIP SN 1999em from outburst to dust formation. *Mon. Not. R. Astron. Soc.* **338**, 939–956 (2003).
24. Chugai, N. N. *et al.* Optical observations of the Type IIP supernova 2004dj: Evidence for the asymmetry of ⁵⁶Ni ejecta. *Astron. Lett.* **31**, 792–805 (2005).
25. Höflich, P. A., Khokhlov, A. & Wang, L. in *20th Texas Symp. on Relativistic Astrophysics* (eds Wheeler, J. C. & Martel, H.) 459–471 (AIP, New York, 2001).
26. Höflich, P. A. Remarks on the polarization observed in SN 1993J. *Astrophys. J.* **440**, 821–829 (1995).
27. Wang, X. *et al.* The progenitor of SN 2004dj in a star cluster. *Astrophys. J.* **626**, L89–L92 (2005).
28. Li, W., Van Dyk, S. D., Filippenko, A. V. & Cuillandre, J.-C. On the progenitor of the Type II SN 2004et in NGC 6946. *Publ. Astron. Soc. Pacif.* **117**, 121–131 (2005).
29. Heger, A., Fryer, C. C., Woosley, S. E., Langer, N. & Hartmann, D. H. How massive stars end their life. *Astrophys. J.* **591**, 288–300 (2003).
30. Maeder, A. & Conti, P. S. Massive star populations in nearby galaxies. *Annu. Rev. Astron. Astrophys.* **32**, 227–275 (1994).

Acknowledgements We thank D. Kasen for discussions. Support for this research was provided by the National Science Foundation, NASA, and the Sylvia and Jim Katzman Foundation. A.V.F. is grateful for a Miller Research Professorship at UC Berkeley, during which part of this work was completed.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.C.L. (leonard@sciences.sdsu.edu).

LETTERS

Channel plasmon subwavelength waveguide components including interferometers and ring resonators

Sergey I. Bozhevolnyi¹, Valentyn S. Volkov¹, Eloïse Devaux², Jean-Yves Laluet² & Thomas W. Ebbesen²

Photonic components are superior to electronic ones in terms of operational bandwidth, but the diffraction limit of light poses a significant challenge to the miniaturization and high-density integration of optical circuits. The main approach to circumvent this problem is to exploit the hybrid nature of surface plasmon polaritons (SPPs), which are light waves coupled to free electron oscillations in a metal^{1,2} that can be laterally confined below the diffraction limit using subwavelength metal structures^{3–8}. However, the simultaneous realization of strong confinement and a propagation loss sufficiently low for practical applications has long been out of reach. Channel SPP modes—channel plasmon polaritons (CPPs)⁸—are electromagnetic waves that are bound to and propagate along the bottom of V-shaped grooves milled in a metal film. They are expected to exhibit useful subwavelength confinement, relatively low propagation loss⁹, single-mode operation¹⁰ and efficient transmission around sharp bends¹¹. Our previous experiments showed that CPPs do exist and that they propagate over tens of micrometres along straight subwavelength grooves¹². Here we report the design, fabrication and characterization of CPP-based subwavelength waveguide components operating at telecom wavelengths: Y-splitters, Mach–Zehnder interferometers and waveguide–ring resonators. We demonstrate that CPP guides can indeed be used for large-angle bending and splitting of radiation, thereby enabling the realization of ultracompact plasmonic components and paving the way for a new class of integrated optical circuits.

We briefly review the main features of CPP propagation along grooves cut into planar metal surfaces^{8–12}. In order to understand the underlying physics of CPP propagation, we first consider SPP modes that are guided inside narrow slits between two metal surfaces, so that the SPPs associated with individual metal surfaces become coupled. The dispersion equation is well known⁷ and its solution can be easily obtained, even when the metal dielectric function is complex. The main feature in this context is that, with the decrease of the slit width, the effective refractive index of a symmetric SPP combination increases (while the propagation length decreases), starting from the value corresponding to the individual SPP^{7,12}. For subwavelength slit widths, the mode field in the slit is nearly constant and close to its maximum, thereby rendering the propagation losses relatively low¹². We now consider a straight groove cut into metal and having a cross-section whose width is monotonously decreasing with depth, for example, a V-shaped groove. As light tends to concentrate in regions with higher refractive indexes, sufficiently deep grooves should support bound symmetric SPP modes that are confined to the groove bottom, where their effective index is at maximum, and predominantly polarized parallel to the metal surface¹². It is important to note that the CPP guiding in V-grooves is counterintuitive: while a certain

groove depth (for a given groove angle) is required to support a CPP mode, the CPP guiding approaches cut-off and the CPP mode field stretches farther out of the groove when the groove angle increases.

Our previous study on CPP guiding¹² allowed us to identify the range of groove parameters for low-loss and well-confined (below the light wavelength) single-mode CPP guiding at telecom wavelengths as well as the corresponding fabrication conditions. In the present work, various waveguide structures consisting of V-grooves with the angles close to 25° and depths of 1.1–1.3 μm have been fabricated (using focused ion-beam milling) in a 1.8- μm -thick gold layer deposited on a substrate of fused silica. The first set of structures to be investigated included straight reference guides, S-bends (that is, curved bends connecting two parallel waveguides offset with respect to each other), Y-splitters composed of two mirrored S-bends, and Mach–Zehnder (MZ) interferometers composed of two consecutive Y-splitters. The design of S-bends was based on sine curves, allowing for continuous curvature throughout the bend¹³. Anticipating (relatively) low loss for S-bends with small curvature radii because of the subwavelength CPP confinement¹², we have chosen the S-bend design that connects two 5- μm -offset waveguides over a distance of 5 μm (the smallest curvature radius is in this case $\sim 2.25 \mu\text{m}$). In general, such a choice is subject to a trade-off between the propagation loss and the bend loss, as the former decreases and the latter increases for shorter bends.

The fabricated plasmonic structures were characterized with a collection scanning near-field optical microscope (SNOM) having an uncoated fibre tip used as a probe and an arrangement for end-fire coupling of tunable (wavelength $\lambda = 1,425\text{--}1,620 \text{ nm}$) TE-polarized (the electric field is parallel to the sample surface plane) radiation into a groove by positioning a tapered-lensed polarization-maintaining single-mode fibre¹². The track of the propagating radiation was clearly distinguishable for all structures and over the whole range of laser tunability. Following the far-field adjustment, the whole fibre–sample arrangement was moved under the SNOM head for near-field mapping of the CPP intensity distribution in the structure under investigation by the uncoated sharp fibre tip of the SNOM. The tip was scanned along the sample surface at a constant distance of a few nanometres maintained by shear force feedback, and the radiation collected by the fibre was detected with a femtowatt InGaAs photoreceiver. Finally, we note that SNOM images are presented here in such a way that the excited CPP propagates in a channel waveguide structure from left to right.

Our SNOM investigations showed that all fabricated structures (S-bends, Y-splitters and MZ interferometers) performed well over the whole range of laser tunability, exhibiting single-mode and subwavelength guiding with overall losses of a few dB, demonstrating thereby robustness to small variations in structural parameters.

¹Department of Physics and Nanotechnology, Aalborg University, Skjernvej 4A, DK-9220 Aalborg Øst, Denmark. ²ISIS, CNRS UMR 7006, Université Louis Pasteur, 8 allée Monge, BP 70028, 67083 Strasbourg, France.

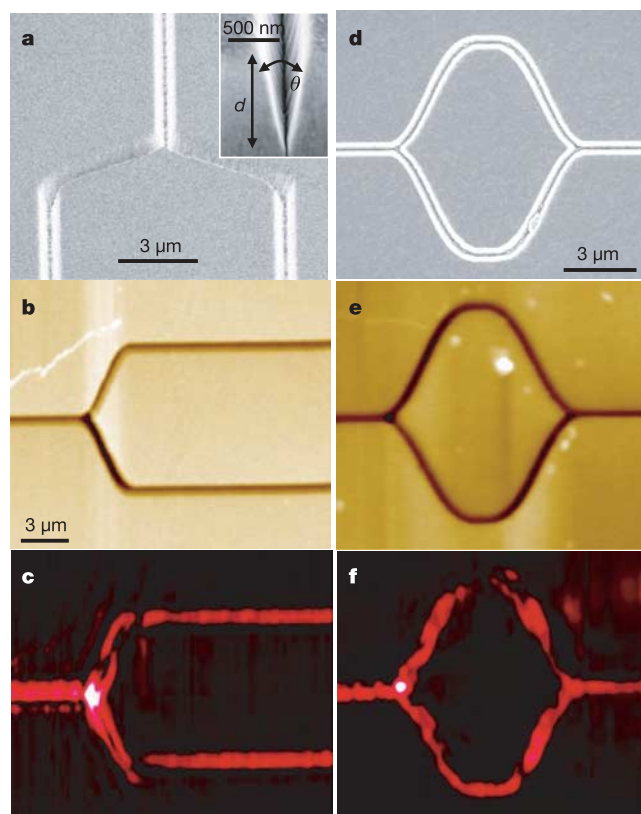


Figure 1 | Plasmonic Y-splitter and Mach-Zehnder (MZ) interferometer. **a**, SEM image, along with **b**, topographical and **c**, near-field optical ($\lambda = 1,600$ nm) SNOM images of the Y-splitter; inset in **a**, SEM image showing typical groove profile (d and θ are the groove depth and angle, respectively). **d–f**, As **a–c** but for the MZ interferometer.

Typical SNOM images of a Y-splitter and an MZ interferometer, along with the corresponding scanning electron microscope (SEM) images, are displayed in Fig. 1. The near-field optical images (Fig. 1c, f) demonstrate very efficient and well-confined CPP guiding throughout the structures consisting of splitters and bends characterized by an average bend angle of 45° . The low loss level observed is a direct (though not straightforward) consequence of the subwavelength CPP field confinement, as discussed earlier¹². Lateral cross-sections of the SNOM images obtained in the course of this investigation showed the average full-width at half-maximum (FWHM) of the CPP mode to be between 0.8 and 1 μm , being almost independent of the wavelength for a given groove—that is, values that are even slightly smaller than those obtained previously¹². An interesting feature observed in most optical images is the signal enhancement in a splitting region (white dot in Fig. 1c, f). We believe that the two main contributions to this effect are the out-of-plane scattering resulting in propagating field components (whose detection efficiency is larger than that of evanescent waves, such as CPPs)¹⁴ and the above-mentioned counterintuitive feature of CPP guiding. In the splitting region, the groove width is nearly doubled, causing the CPP guiding to approach cut-off, the effective centre of CPP mode field to move upwards (out of the groove) and, thereby, the detected optical signal to increase.

Quantitative evaluation of the insertion loss was made by considering average cross-sections before and after the component in question, that is, Y-splitters or MZ interferometers (Fig. 2). These cross-sections were obtained with the optical images shown in Fig. 1 by averaging within $\sim 3\text{-}\mu\text{m}$ -long windows placed in straight groove regions. It is seen that the optical signal outside the grooves amounts to $\sim 15\%$ of the maximum signal associated with the CPP mode at the input regions. This background signal becomes the main

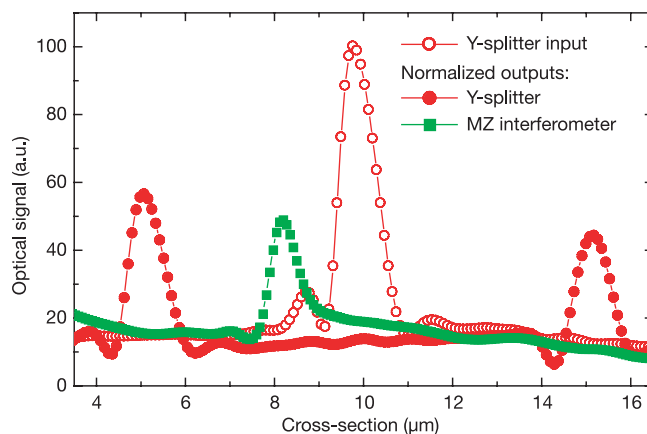


Figure 2 | Input and output mode profiles. Normalized cross-sections obtained with the optical images shown in Fig. 1c and f by using $\sim 3\text{-}\mu\text{m}$ -long averaging windows along straight groove regions. The optical signal detected with the MZ interferometer was normalized to be in the input channel of the Y-splitter. a.u., arbitrary units.

contribution in the optical images recorded in constant plane mode at the height of ~ 200 nm above the sample surface¹². However, it is not clear to what extent it influences the signal measured in the middle of a groove with the SNOM fibre tip being actually below the sample surface by $\sim 40\text{--}100$ nm, depending on the groove and tip shapes. With this in mind, one obtains for the transmission T of the Y-splitter and MZ interferometer the following ranges: $0.82 < T_Y < 1$ and $0.4 < T_{MZ} < 0.49$, respectively, using the output-to-input ratios of maxima of the corresponding signal distributions (Fig. 2). Similar values were found for the Y-splitter and MZ-interferometer transmissions at other wavelengths exhibiting (unsystematic) variations within 20% that can probably be attributed to different background levels (determined by positioning of the in-coupling fibre with respect to the input groove).

The fact that the investigated Y-splitter showed next to zero insertion loss is remarkable, considering the fact that the smallest curvature radius used was only $\sim 1.5\lambda$. This demonstrates the potential of using similar groove structures in sophisticated plasmonic circuits, as well as the practical feasibility of their fabrication with sufficiently high quality. Our investigations have also revealed that these structures exhibit rather high tolerance to topological imperfections on the nanometre scale, that is, there seems to be little perturbation due to accidental presence of nanoparticles (for example, one particle sitting in the lower arm of MZ interferometer is clearly seen in Fig. 1d). Note that, on the other hand, the choice of the groove angle is subject to trade-off, as better CPP confinement achieved with narrower grooves would ensure smaller bend losses, causing at the same time larger propagation losses¹².

The results obtained with the above set of structures encouraged us to further exploit the potential of CPPs and to realize a functional plasmonic component. As an example, we have chosen a waveguide–ring (WR) resonator, which is interesting owing to its wavelength response and non-trivial dependence on the coupling between straight and ring waveguides¹⁵. The main system parameters are transmission and coupling coefficients, t and κ , and transmission around the ring, $\alpha \exp(i\theta)$ (inset in Fig. 3a). The transmission through a WR resonator under the condition that a single unidirectional mode of the ring resonator is excited is given by¹⁵

$$T = \frac{\alpha^2 \eta^2 + |t|^2 - 2\alpha\eta|t|\cos(\theta + \phi)}{1 + \alpha^2|t|^2 - 2\alpha|t|\cos(\theta + \phi)}, \quad (1)$$

$$t = |t|\exp(i\phi), |t|^2 + |\kappa|^2 = \eta, \theta = \frac{2\pi}{\lambda} N_{\text{eff}} 2\pi R$$

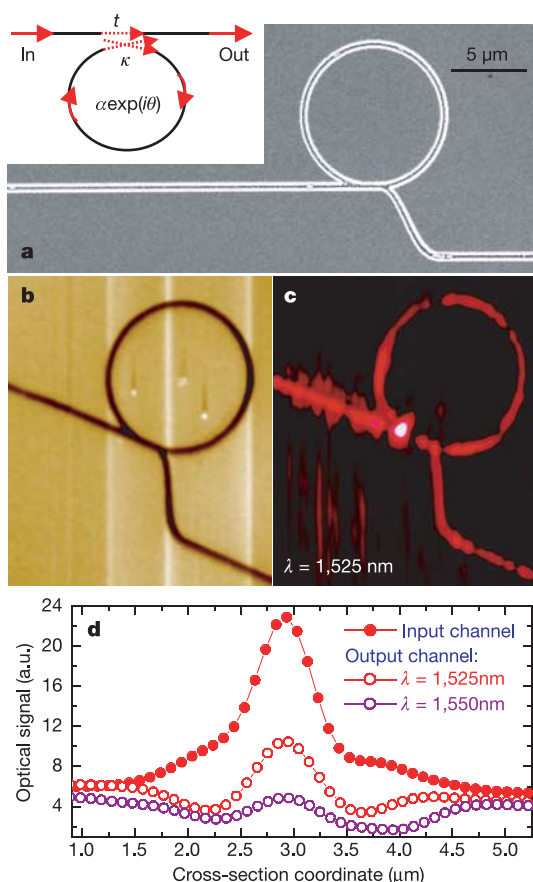


Figure 3 | Plasmonic waveguide-ring (WR) resonator. **a**, SEM image, along with **b**, topographical and **c**, near-field optical (wavelength $\lambda = 1,525$ nm) SNOM images of the WR resonator schematically shown in the inset in **a** (t and κ are the transmission and coupling coefficients, respectively, and $\alpha\exp(i\theta)$ denotes the amplitude transmission around the ring). **d**, Normalized cross-sections obtained with the optical images (one is shown in **c**) recorded for two different wavelengths with the signal being normalized to be at the same level in the input channel.

where η is the overall coupling efficiency, which is influenced by the radiation (power) loss at the coupling region, R is the ring radius, N_{eff} is the CPP effective index, and λ is the light wavelength in air. It is seen that the transmission is periodic with respect to the phase θ accumulated by the CPP mode per circulation, implying the possibility of wavelength filtering. At resonance ($\theta + \phi = 2\pi m$, m is integer), the transmitted power vanishes at the condition of critical coupling, $\alpha\eta = |t|$, allowing for dramatic influence of the ring transmission α on the WR resonator transmission and other interesting effects¹⁵.

Our main interest was in verifying the possibility of realization of a plasmonic WR resonator and achieving thereby the wavelength dependent transmission of CPPs. In the design of WR resonator, we used a 5- μm -radius ring and the same S-bend as in the first set of structures (Fig. 3a). The intention with using an S-bend after the coupling region and laterally displacing the output channel was to decrease the influence of radiation scattered in the coupling region on the signal detected in the output channel. The fabricated groove structure was 1.3- μm deep with a groove angle of $\sim 25^\circ$. We have already noticed during the far-field adjustment that, contrary to the previously investigated structures, the transmitted power depended on the wavelength of light used. We then recorded SNOM images for different wavelengths in the laser tunability range for two nominally identical WR resonators. A typical optical image obtained with the transmitted light power at maximum is shown in Fig. 3c, along with the corresponding input and output average cross-sections for two different wavelengths exhibiting a clear difference

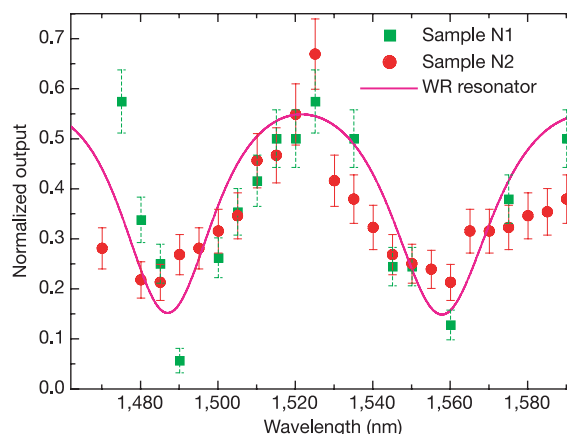


Figure 4 | Transmission spectra for two WR resonators. The output-to-input ratio obtained by using average cross-sections of optical images (similar to that shown in Fig. 3d) as a function of the light wavelength for two nominally identical WR resonators (one is shown in Fig. 3a), along with the fitted curve corresponding to the response of the WR resonator (see text). The error bars are estimated from the maximum variation in the ratio when placing the averaging windows at different positions along the input and output channels.

in the transmitted CPP power (Fig. 3d). Similar cross-sections were used in order to evaluate the normalized output, that is, transmission, for the investigated WR resonators as a function of the light wavelength.

The data obtained for both structures were fitted with the dependence calculated according to equation (1), demonstrating quite good agreement (Fig. 4). When fitting, we used the CPP parameters as previously calculated¹² ($N_{\text{eff}} = 1.015$, $\alpha = 0.77$, obtained from the calculated CPP propagation length and ring circumference, neglecting the bend loss), assumed the transmission coefficient to be real, that is, $\phi = 0$, and allowed for small variations in the ring radius R . The fitted values were found to be $\eta \approx 0.78$, $t \approx 0.33$ and $R \approx 5.13 \mu\text{m}$, which are consistent with the previous experiments (for example, the overall coupling efficiency η is relatively high, implying low radiation loss, similar to that seen above with Y-splitters and MZ interferometers). Note that the achieved maximum-to-minimum transmission ratio is quite large (on average, >3) allowing for wavelength filtering with a bandwidth of ~ 40 nm.

It should be stressed that, in contrast to the CPP-based structures, the degree of light confinement in dielectric waveguides, including those based on the photonic bandgap effect, is fundamentally limited by the light wavelength in the dielectric used. Here, we have demonstrated a WR resonator with an insertion loss of <3 dB and a footprint of only $200 \mu\text{m}^2$ ($\sim 100 \lambda^2$) that can be used for wavelength filtering with a bandwidth of ~ 40 nm. The performance of the considered structures could be further improved by perfecting the quality of fabricated grooves (especially walls and junctions) and optimizing geometrical parameters with respect to the above-mentioned trade-off. We can envisage further development of this concept for miniature bio-sensors and for ultracompact plasmonic interconnects that can be naturally integrated with electrical circuits in a way similar to that recently demonstrated with long-range SPPs¹⁶.

Received 7 November 2005; accepted 18 January 2006.

1. Raether, H. *Surface Plasmons* (Springer, Berlin, 1988).
2. Barnes, W. L., Dereux, A. & Ebbesen, T. W. Surface plasmon subwavelength optics. *Nature* **424**, 824–830 (2003).
3. Weeber, J. C. et al. Near-field observation of surface plasmon polariton propagation on thin metal stripes. *Phys. Rev. B* **64**, 045411 (2001).
4. Takahara, J., Yamagishi, S., Taki, H., Morimoto, A. & Kobayashi, T. Guiding of a one-dimensional optical beam with nanometer diameter. *Opt. Lett.* **22**, 475–477 (1997).

5. Quinten, M., Leitner, A., Krenn, J. R. & Aussenegg, F. R. Electromagnetic energy transport via linear chains of silver nanoparticles. *Opt. Lett.* **23**, 1331–1333 (1998).
6. Maier, S. A. *et al.* Local detection of electromagnetic energy transport below the diffraction limit in metal nanoparticle plasmon waveguides. *Nature Mater.* **2**, 229–232 (2003).
7. Tanaka, K. & Tanaka, M. Simulations of nanometric optical circuits based on surface plasmon polariton gap waveguide. *Appl. Phys. Lett.* **82**, 1158–1160 (2003).
8. Novikov, I. V. & Maradudin, A. A. Channel polaritons. *Phys. Rev. B* **66**, 035403 (2002).
9. Pile, D. F. P. & Gramotnev, D. K. Channel plasmon-polariton in a triangular groove on a metal surface. *Opt. Lett.* **29**, 1069–1071 (2004).
10. Gramotnev, D. K. & Pile, D. F. P. Single-mode subwavelength waveguide with channel plasmon-polaritons in triangular grooves on a metal surface. *Appl. Phys. Lett.* **85**, 6323–6325 (2004).
11. Pile, D. F. P. & Gramotnev, D. K. Plasmonic subwavelength waveguides: next to zero losses at sharp bends. *Opt. Lett.* **30**, 1186–1188 (2005).
12. Bozhevolnyi, S. I., Volkov, V. S., Devaux, E. & Ebbesen, T. W. Channel plasmon-polariton guiding by subwavelength metal grooves. *Phys. Rev. Lett.* **95**, 046802 (2005).
13. Kumar, A. & Aditya, S. Performance of S-bends for integrated-optic waveguides. *Microw. Opt. Technol. Lett.* **19**, 289–292 (1998).
14. Bozhevolnyi, S. I. & Volkov, V. S. Near-field characterization of planar photonic-crystal-waveguide structures. *Phil. Trans. R. Soc. Lond. A* **362**, 757–769 (2004).
15. Yariv, A. Universal relations for coupling of optical power between microresonators and dielectric waveguides. *Electron. Lett.* **36**, 321–322 (2000).
16. Nikolajsen, T., Leosson, K. & Bozhevolnyi, S. I. Surface plasmon polariton based modulators and switches operating at telecom wavelengths. *Appl. Phys. Lett.* **85**, 5833–5835 (2004).

Acknowledgements This work was supported by the European Network of Excellence, PLASMO-NANO-DEVICES and STREP SPP.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.I.B. (sergey@physics.aau.dk).

LETTERS

Mechanical twisting of a guest by a photoresponsive host

Takahiro Muraoka^{1*}, Kazushi Kinbara^{1,2*} & Takuzo Aida^{1*}

Molecular analogues of a variety of mechanical devices such as shuttles, brakes, unidirectional rotors and tweezers have been created^{1–11}. But these ‘molecular machines’ have not yet been used to mechanically manipulate a second molecule in a controlled and reversible manner. Here we show that light-induced scissor-like conformational changes of one molecule⁵ can give rise to mechanical twisting of a non-covalently bound guest molecule. To realize this coupling of molecular motions, we use a previously designed system⁵: a ferrocene moiety with an azobenzene strap, each end of which is attached to one of the two cyclopentadienyl rings of the ferrocene unit, acts as a pivot so that photoisomerization of the strap rotates the ferrocene rings relative to each other and thereby also changes the relative position of two ‘pedal’ moieties attached to the ferrocene rings. We translate this effect into intermolecular coupling of motion by endowing the pedals with binding sites, which allow the host system to form a stable complex with a bidentate rotor molecule. Using circular dichroism spectroscopy, we show that the photoinduced conformational changes of the host are indeed transmitted and induce mechanical twisting of the rotor molecule. This design concept, which significantly extends the successful coupling of motion beyond the intramolecular level seen in synthetic allosteric receptors¹², might allow for the remote control of molecular events in larger interlocked molecular systems.

The design of our molecular system is illustrated in Fig. 1. As shown, the host molecule (1) bears a zinc porphyrin unit at each cyclopentadienyl ring of the ferrocene unit. Since zinc porphyrins are known to coordinate to nitrogenous bases, compound 1 is expected to bind the bidentate molecular rotor (2) to form a stable host/guest complex 2·1. In this system, photoisomerization of the azobenzene strap in response to irradiation with ultraviolet (UV) and visible light should induce a scissor- or pedal-like conformational change of the zinc porphyrin units, which can be used to twist the rotary guest 2 repeatedly in clockwise and counterclockwise directions. The synthesis of compound 1 with a *trans* configuration at the azobenzene unit (*trans*-1) proceeded according to the method described in the Supplementary Information.

When characterized by circular dichroism (CD) spectroscopy, the enantiomers of *trans*-1 showed characteristic mirror-image CD bands of one another at the absorption bands of the ferrocene (230–370 nm) and zinc porphyrin units (380–440 nm) (Supplementary Fig. S2), suggesting a chirally twisted geometry of the zinc porphyrin units. Exposure of *trans*-1 to UV light ($\lambda = 350 \pm 10$ nm) at 20 °C induces *trans*-to-*cis* isomerization of the azobenzene unit to give a molar ratio for *trans*-1/*cis*-1 of 22/78 (Supplementary Figs S3a–c and Fig. S4a). Irradiation of the isomerized mixture with visible light ($\lambda > 420$ nm) resulted in the backward *cis*-to-*trans* isomerization, giving a molar ratio for *trans*-1/*cis*-1 of 63/37

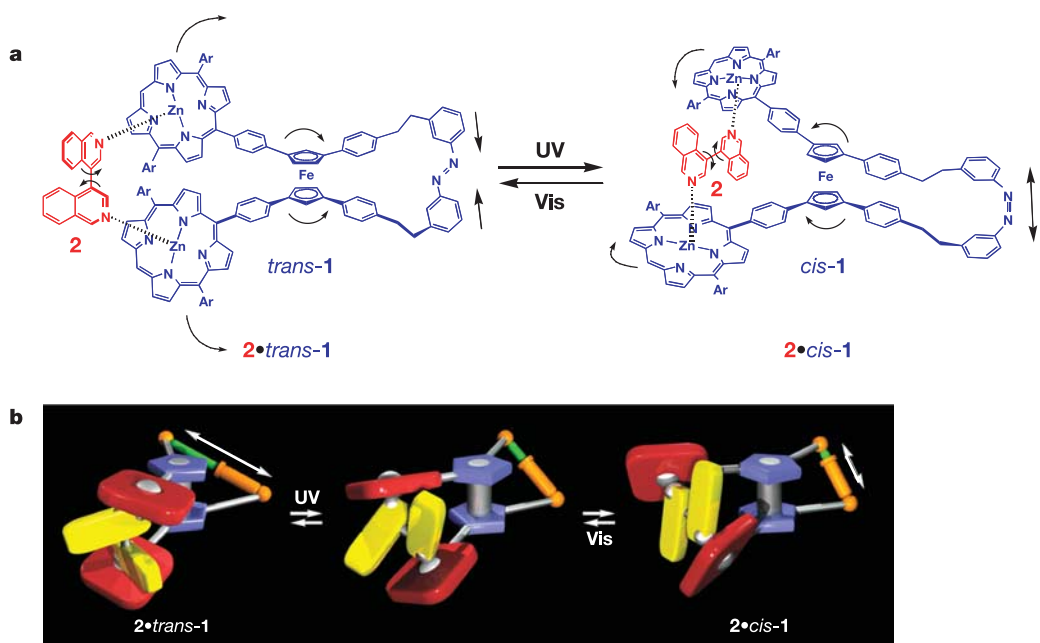


Figure 1 | Design and concept of light-powered molecular pedal.

a, Schematic representation of photoisomerization of a 1:1 complex of molecular pedal 1 with rotary guest 2 (2·1). Ar, 4-decyloxyphenyl. Arrows indicate the directions of interlocked motions. **b**, Schematic illustrations of a sequence of interlocked motions of 2·1 triggered by light. Orange/green, blue and red-coloured parts show the azobenzene, ferrocene and zinc porphyrin units of 1, respectively, while the yellow part denotes 2.

¹Department of Chemistry and Biotechnology, School of Engineering, and Centre for NanoBio Integration, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan. ²PRESTO, Japan Science Technology Agency (JST), 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan.

*These authors contributed equally to this work.

(Supplementary Figs S3d–f and S4b). The photoisomerization events were accompanied by a notable CD spectral change of **1** both at the ferrocene and zinc porphyrin units (Supplementary Fig. S2). These results indicate that the contraction and elongation of the azobenzene strap are translated into clockwise and counterclockwise rotary motions of the cyclopentadienyl rings, leading to a pedal-like motion of the guest-binding zinc porphyrin units. ^1H NMR spectroscopy provided further evidence for these interlocked motions (Supplementary Fig. S5).

Compound **2** is expected to behave as a bidentate ligand and to interact strongly with **1**. Indeed, addition of **2** in CH_2Cl_2 resulted in *trans*-**1** displaying red-shifted Soret and Q bands of the zinc porphyrin units with isosbestic points at 420 and 554 nm, respectively (Fig. 2a)^{13–15}. The spectroscopic data and Job's plot indicate that *trans*-**1** binds only one molecule of **2** to form **2:trans**-**1** with an association constant K_{assoc} of $1.3 \times 10^6 \text{ M}^{-1}$ (filled squares in Supplementary Fig. S6, and Fig. S7a). This large K_{assoc} value is obviously due to the complexation involving two contact points (Fig. 1a), which is not only important to achieve tight host–guest docking but also essential for the transmission of motion from **1** to **2**.

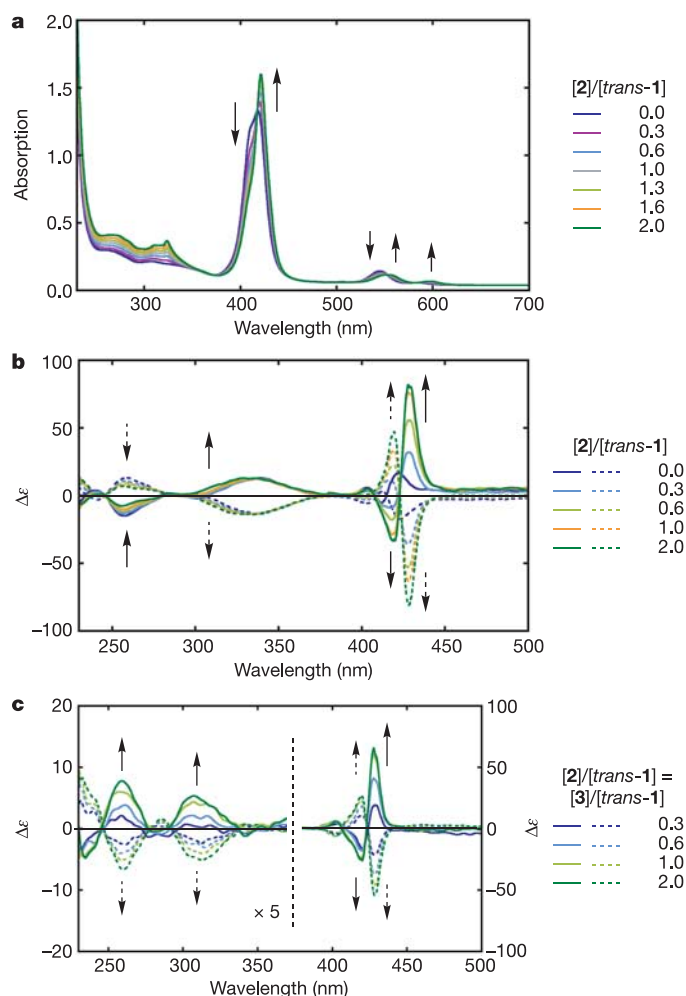


Figure 2 | Binding of rotary guest **2 with molecular pedal *trans*-**1**.** Absorption (a) and CD (b) spectral changes of *trans*-**1** in CH_2Cl_2 ($1.1 \times 10^{-5} \text{ M}$) at 20 °C upon titration with **2**. In a, the [2]/[*trans*-**1**] ratio was 0.0, 0.3, 0.6, 1.0, 1.3, 1.6 and 2.0; in b, [2]/[*trans*-**1**] ratio was 0.0, 0.3, 0.6, 1.0 and 2.0. c, Differential spectra by subtraction of the CD spectra of **3:trans**-**1** from those of **2:trans**-**1** at [2]/[*trans*-**1**] = [3]/[*trans*-**1**] = 0.3, 0.6, 1.0 and 2.0. Solid curves, (+)-*trans*-**1**; broken curves, (–)-*trans*-**1**. Arrows indicate the directions of spectral changes. $\Delta\epsilon$ represents molar circular dichroic absorption coefficient ($\text{L mol}^{-1} \text{ cm}^{-1}$).

When a CH_2Cl_2 solution of **2:trans**-**1** was irradiated with UV light, isomerization of *trans*-**1** occurred to furnish a molar ratio *trans*-**1**/*cis*-**1** of 22/78 (Supplementary Fig. S8). Because *cis*-**1** was hardly isolable due to its high susceptibility to the backward isomerization, the stoichiometry of the complexation of *cis*-**1** with **2** was evaluated based on spectral changes upon UV irradiation of the Job's plot samples used for the *trans*-**1** case (see above). The maximum observed at the 1:1 molar ratio of *trans*-**1** to **2** remained unchanged after the UV irradiation (filled circles in Supplementary Fig. S6),

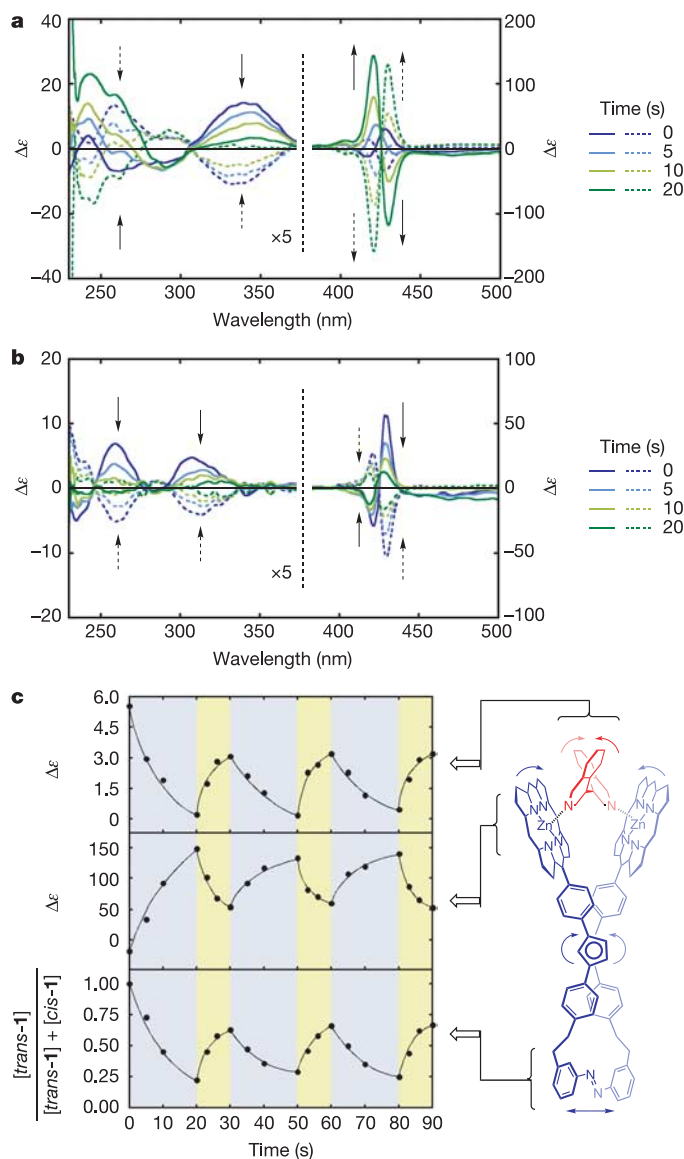


Figure 3 | Circular dichroism (CD) visualization of the motions of guest-binding molecular pedals **2-(+)-**1** and **2**-(-)-**1** triggered by light.** a, CD spectral changes of **2:trans**-**1** in CH_2Cl_2 at 20 °C upon UV irradiation ($\lambda = 350 \pm 10 \text{ nm}$) for 5, 10 and 20 s. [*trans*-**1**] = $7.2 \times 10^{-6} \text{ M}$; [2]/[*trans*-**1**] = 1.0. Solid curves, (+)-*trans*-**1**; broken curves, (–)-*trans*-**1**. b, Differential spectra by subtraction of the CD spectra of **3:trans**-**1** from those of **2:trans**-**1** in CH_2Cl_2 at 20 °C upon UV irradiation ($\lambda = 350 \pm 10 \text{ nm}$) for 5, 10 and 20 s. [*trans*-**1**] = $7.2 \times 10^{-6} \text{ M}$; [2]/[*trans*-**1**] = [3]/[*trans*-**1**] = 1.0. Solid curves, (+)-*trans*-**1**; broken curves, (–)-*trans*-**1**. c, CD intensity changes of **2**-(+)-*trans*-**1** at 307.4 nm (top) in Fig. 3b and 420.8 nm (middle) in Fig. 3a, along with a change in mole fraction of *trans*-**1** in **2**:**1** (bottom), upon alternating irradiation of UV ($\lambda = 350 \pm 10 \text{ nm}$, blue regions) and visible lights ($\lambda > 420 \text{ nm}$, light-green regions). Arrows in a and b indicate the directions of spectral changes. Arrows in c represent the directions of interlocked motions.

indicating that *cis*-1 as well as *trans*-1 forms a 1:1 complex with 2 (**2**·*cis*-1). On the basis of this result, we titrated a *cis*-enriched isomeric mixture of 1 (*trans*-1/*cis*-1 = 22/78) with 2 and estimated K_{assoc} for the complexation between *cis*-1 and 2 as $2.6 \times 10^6 \text{ M}^{-1}$, which is as large as that of *trans*-1 with 2 (Supplementary Fig. S7b).

We conducted CD spectroscopic studies on the photoisomerization of the host/guest complex using (+)-*trans*-1, an enantiomer of *trans*-1 exhibiting a positive CD band at 421 nm. When (+)-*trans*-1 was complexed with guest 2 in CH_2Cl_2 , new CD bands appeared at around the red-shifted Soret absorption band of the guest-bound zinc porphyrin units, at the expense of the original Cotton effects at 421 nm due to guest-free (+)-*trans*-1, and the spectral change finally reached a plateau at a molar ratio for **2**/1 of 2.0 (Fig. 2b). As expected, titration of (–)-*trans*-1, an enantiomeric counterpart of (+)-*trans*-1, with 2 resulted in a mirror-image CD spectral change of (+)-*trans*-1. Since the CD activity of **2**·(+)-*trans*-1 is much larger than that of guest-free (+)-*trans*-1, the twisted geometry of the two zinc porphyrin units in *trans*-1 is fixed upon complexation with bidentate guest 2. The exciton chirality method^{14,15} indicated that the zinc porphyrin units in **2**·(+)-*trans*-1 and **2**·(–)-*trans*-1 might adopt *P* and *M* configurations, respectively.

Interestingly, when *trans*-1 in these host–guest complexes was isomerized to the *cis* form by UV irradiation, a large CD spectral change, with an inversion of its sign at the final stage, took place at the Soret absorption band of the zinc porphyrin units (Fig. 3a). Upon exposure of the resulting complex to visible light to allow the backward isomerization, an inverse CD spectral change took place. The large spectral changes accompanying the photoisomerization of **2**·1 indicate that the light-induced isomerization that contracts and elongates the azobenzene strap gives rise to a scissor- or pedal-like motion of the zinc porphyrin units even when they bind a guest molecule.

In view of this behaviour and the large association constants of rotary guest 2 with *trans*- and *cis*-1, we expected that 2 could be twisted by 1. This can be probed by considering that rotary guest 2 will be optically active if it adopts a nonplanar, chiral geometry. Because of a freedom in conformational change, 2 is not optically active in solution but turns optically active upon binding with 1, with the CD activity of the latter changing in response to the

photoisomerization of the azobenzene unit. Chiral compound 2 is expected to give CD bands at 270–350 nm, considering its absorption spectral profile. However, the CD spectrum of **2**·1 in this region involves Cotton effects of the host component itself (1). Therefore, to selectively visualize the conformational motion of 2, the contribution of 1 must be deducted from the overall CD activity of **2**·1.

For evaluating the net contribution of 1 to the CD profile of **2**·1 at 270–350 nm, we used 3,3'-bipyridine (**3**) as an alternative guest, which is spectroscopically transparent and should therefore be CD-silent in this region even though it adopts a chiral geometry. We confirmed that compound 3 forms a highly stable 1:1 complex with (+)-*trans*-1 ($K_{\text{assoc}} = 1.2 \times 10^6 \text{ M}^{-1}$), displaying a CD titration profile analogous to the case of 2 (Supplementary Figs S7c and S9). Figure 2c shows differential CD spectra between **2**·(+)-*trans*-1 and **3**·(+)-*trans*-1 at varying guest concentrations, which clearly display two positive Cotton effects at 259 and 307 nm. When (–)-*trans*-1 was used instead of (+)-*trans*-1 for the complexation with 2, inversely signed Cotton effects resulted. Because the differential CD spectra are quite analogous to the CD spectral profiles of the enantiomers of 4,4'-di-(3-methylisoquinoline) (**4**) that are separable due to a restricted conformational change (Supplementary Fig. S10), we conclude that rotary guest 2, upon complexation with 1, is frozen in an optically active conformation owing to chirality transfer from the host component. By reference to the signs of the Cotton effects of the enantiomers of 4, the rotary guest 2 in **2**·(+)-*trans*-1 and **2**·(–)-*trans*-1 most probably adopts (*R*)- and (*S*)-configurations, respectively. As illustrated in Fig. 4, these geometries are complementary in shape to the *P* and *M* configurations suggested for the corresponding zinc porphyrin units.

We next investigated changes in the CD spectral profiles of rotary guest 2 in **2**·(+)-*trans*-1 and **2**·(–)-*trans*-1 in response to the photoisomerization of the azobenzene strap (Fig. 3b). Upon irradiation with UV light to induce the *trans*-to-*cis* isomerization of the strap, the Cotton effects at 270–350 nm due to 2 became smaller and almost disappeared. On the other hand, when the resulting complex was irradiated with visible light to allow its backward isomerization, these CD bands appeared again and grew up with time. Clearly, the CD spectral change of **2** (Fig. 3c, top) upon photoisomerization (Fig. 3c, bottom) was synchronous to that observed for the zinc porphyrin units of 1 (Fig. 3c, middle), demonstrating that non-covalently docked host molecule 1 and rotary guest 2 are stereochemically interlocked with one another. The observed optical activities of rotary guest 2 upon binding with *trans*- and *cis*-1 are consistent with those expected from computer-generated molecular structures of **2**·*trans*-1 and **2**·*cis*-1, where the two isoquinoline rings of 2 docked at *trans*-1 adopt a twisted geometry with one another, while those docked at *cis*-1 adopt a nearly planar geometry with a lower CD activity^{16,17} (Supplementary Video 1).

The above observations clearly establish that the rotary guest 2 adjusts its conformation in response to conformational changes in 1. But in principle, this response might proceed through dissociation of 2 from 1, and rebinding of guest 2 only after 1 has changed its structure. However, a consideration of relevant timescales argues against such a process. The photoisomerization of azobenzene derivatives is known to occur on a timescale of subpicoseconds to picoseconds¹⁸. Systems relevant to the present study are azobenzene-containing cyclic oligopeptides, where photoisomerization of the azobenzene unit in a constrained environment takes place in picoseconds and is followed by a conformational change of the connecting peptide units on a timescale of picoseconds to nanoseconds^{19,20}. Moreover, the cyclopentadienyl rings in ferrocene derivatives are known to rotate on a timescale of nanoseconds²¹.

Thus, the ferrocene unit in 1 can respond rapidly through rotation to the photoisomerization of the azobenzene unit and the subsequent conformational changes of the connecting units, and thereby induce the guest-binding zinc porphyrin units to undergo a pedal- or scissor-like motion. For this motion to be directly transmitted to

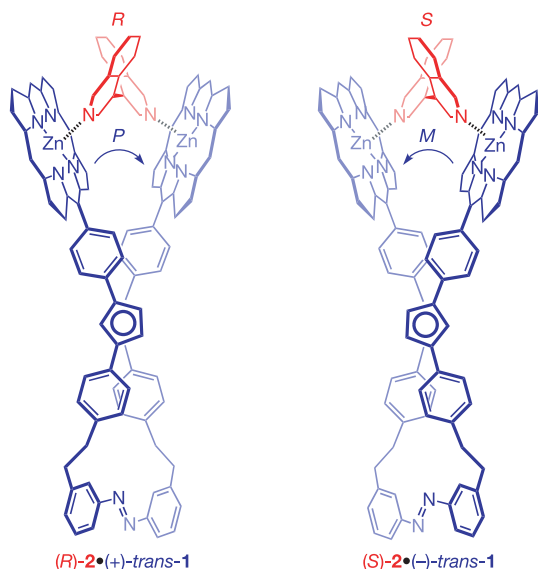


Figure 4 | Configurational aspects of the zinc porphyrin units of molecular pedal *trans*-1 and rotary guest 2 in **2**·*trans*-1. The configurations of the zinc porphyrin units are estimated using the exciton chirality method^{15,16}, while those of 2 in **2**·(+)-*trans*-1 and **2**·(–)-*trans*-1 are deduced by analogy with the CD spectra of (*R*)-**4** and (*S*)-**4**, respectively.

the rotary guest (2), the dissociation dynamics of 2·1 must occur on a much slower timescale than the motion itself. We therefore conducted a ¹H NMR lineshape analysis of the binding behaviour of 2·*trans*-1 and 2·*cis*-1 using program DNMR5²² and found that guest dissociation occurs only very slowly, on a timescale of 0.6 and 0.2 ms, respectively (Supplementary Fig. S11). Guest dissociation thus occurs at a rate that is at least six orders of magnitude slower than the photoisomerization of azobenzene derivatives, which provides compelling evidence that rotary guest 2 will essentially remain associated with its host 1 throughout the latter's photoinduced conformational change and as a result be physically twisted.

Received 30 November 2005; accepted 3 February 2006.

- Kinbara, K. & Aida, T. Toward intelligent molecular machines: directed motions of biological and artificial molecules and assemblies. *Chem. Rev.* **105**, 1377–1400 (2005).
- Collin, J.-P., Dietrich-Buchecker, C., Gaviña, P., Jimenez-Molero, M. C. & Sauvage, J.-P. Shuttles and muscles: linear molecular machines based on transition metals. *Acc. Chem. Res.* **34**, 477–487 (2001).
- Balzani, V., Credi, A., Raymo, F. M. & Stoddart, J. F. Artificial molecular machines. *Angew. Chem. Int. Edn Engl.* **39**, 3348–3391 (2000).
- Badjic, J. D., Balzani, V., Credi, A., Silvi, S. & Stoddart, J. F. A molecular elevator. *Science* **303**, 1845–1849 (2004).
- Muraoka, T., Kinbara, K., Kobayashi, Y. & Aida, T. Light-driven open–close motion of chiral molecular scissors. *J. Am. Chem. Soc.* **125**, 5612–5613 (2003).
- Ishii, D. *et al.* Chaperonin-mediated stabilization and ATP-triggered release of semiconductor nanoparticles. *Nature* **423**, 628–632 (2003).
- Leigh, D. A., Wong, J. K. Y., Dehez, F. & Zerbetto, F. Unidirectional rotation in a mechanically interlocked molecular rotor. *Nature* **424**, 174–179 (2003).
- Tashiro, K., Konishi, K. & Aida, T. Metal bisporphyrinate double-decker complexes as redox-responsive rotating modules. Studies on ligand rotation activities of the reduced and oxidized forms using chirality as a probe. *J. Am. Chem. Soc.* **122**, 7921–7926 (2000).
- Kelly, T. R., De Silva, H. & Silva, R. A. Unidirectional rotary motion in a molecular system. *Nature* **401**, 150–152 (1999).
- Koumura, N., Zijlstra, R. W. J., van Delden, R. A., Harada, N. & Feringa, B. L. Light-driven monodirectional molecular rotor. *Nature* **401**, 152–155 (1999).
- Shinkai, S., Ogawa, T., Nakaji, T., Kusano, Y. & Manabe, O. Photocontrolled extraction ability of azobenzene-bridged azacrown ether. *Tetrahedr. Lett.* **20**, 4569–4572 (1979).
- Rebek, J. & Wattle, R. V. Allosteric effects: remote control of ion transport selectivity. *J. Am. Chem. Soc.* **102**, 4853–4854 (1980).
- Guo, Y.-M., Oike, H. & Aida, T. Chiroptical transcription of helical information through supramolecular harmonization with dynamic helices. *J. Am. Chem. Soc.* **126**, 716–717 (2004).
- Borovkov, V. V., Hembury, G. A. & Inoue, Y. Origin, control, and application of supramolecular chirogenesis in bisporphyrin-based systems. *Acc. Chem. Res.* **37**, 449–459 (2004).
- Huang, X. *et al.* Absolute configurational assignments of secondary amines by CD-sensitive dimeric zinc porphyrin host. *J. Am. Chem. Soc.* **124**, 10320–10335 (2002).
- Harada, N., Chen, S. L. & Nakanishi, K. Quantitative definition of exciton chirality and the distant effect in the exciton chirality method. *J. Am. Chem. Soc.* **97**, 5345–5352 (1975).
- Hanazaki, I. & Akimoto, H. Optical rotatory power of 2,2'-dihydroxy-1,1'-binaphthyl and related compounds. *J. Am. Chem. Soc.* **94**, 4102–4106 (1972).
- Tamai, N. & Miyasaka, H. Ultrafast dynamics of photochromic system. *Chem. Rev.* **100**, 1875–1890 (2000).
- Spörlein, S. *et al.* Ultrafast spectroscopy reveals subnanosecond peptide conformational dynamics and validates molecular dynamics simulation. *Proc. Natl Acad. Sci. USA* **99**, 7998–8002 (2002).
- Wachtveitl, J. *et al.* Ultrafast conformational dynamics in cyclic azobenzene peptides of increased flexibility. *Biophys. J.* **86**, 2350–2362 (2004).
- Gardner, A. B., Howard, J., Waddington, T. C., Richardson, R. M. & Tomkinson, J. The dynamics of ring rotation in ferrocene, nickelocene, and ruthenocene by incoherent quasi-elastic neutron scattering. *Chem. Phys.* **57**, 453–460 (1981).
- Stephenson, D. S. & Binsch, G. Improved algorithm for the computation of exchange-broadened NMR bandshapes. *J. Magn. Reson.* **30**, 625–626 (1978).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements T.M. thanks the JSPS Young Scientist Fellowship. This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research on Priority Areas 'Life-Surveyors' (to K.K.). We thank Y. Casta and S. Nara for cooperation in synthesis, and N. Tamai and M. Irie for discussion on photophysics.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.K. (kinbara@macro.t.u-tokyo.ac.jp) and T.A. (aida@macro.t.u-tokyo.ac.jp).

LETTERS

Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era

Yuichiro Ueno^{1,3,5}, Keita Yamada^{4,5}, Naohiro Yoshida^{1,3,4,5}, Shigenori Maruyama^{1,2} & Yukio Isozaki⁶

Methanogenic microbes may be one of the most primitive organisms¹, although it is uncertain when methanogens first appeared on Earth. During the Archaean era (before 2.5 Gyr ago), methanogens may have been important in regulating climate, because they could have provided sufficient amounts of the greenhouse gas methane to mitigate a severely frozen condition that could have resulted from lower solar luminosity² during these times. Nevertheless, no direct geological evidence has hitherto been available in support of the existence of methanogens in the Archaean period, although circumstantial evidence is available in the form of ~2.8-Gyr-old carbon-isotope-depleted kerogen³. Here we report crushing extraction and carbon isotope analysis of methane-bearing fluid inclusions in ~3.5-Gyr-old hydrothermal precipitates from Pilbara craton, Australia. Our results indicate that the extracted fluids contain microbial methane with carbon isotopic compositions of less than -56‰ included within original precipitates. This provides the oldest evidence of methanogen (>3.46 Gyr ago), pre-dating previous geochemical evidence by about 700 million years.

The samples came from the Dresser Formation at the North Pole area in Pilbara craton, Western Australia. The Dresser Formation consists of pillowed basaltic greenstones and chert beds. In the North Pole area these rocks have merely undergone low-grade metamorphism below the greenschist facies⁴⁻⁶. The lowermost chert unit of the formation is intercalated with several barite beds⁴. This chert-barite unit contains ³⁴S-depleted pyrites, which were presumably produced by sulphate-reducing microbes⁷. The unit also contains one of the oldest putative microfossils^{8,9} (see refs 10 and 11 for alternative views).

The minimum depositional age of the Dresser Formation is constrained by the zircon U-Pb age of $3,458 \pm 2$ Myr for the felsic volcanics that overlie the formation¹². A model lead age of 3,490 Myr¹² was obtained for galena from the chert-barite unit. This may represent the actual depositional age of the Dresser Formation^{13,14}.

In the study area, more than 2,000 silica dykes characteristically intruded into the pillowed basaltic greenstones below the chert-barite unit^{5,6,9,13,14} (Fig. 1 and Supplementary Fig. S1). These are 0.3–20 m wide and generally more than 100 m in length; the longest is more than 1 km in length. Some silica dykes intruded into the chert beds of the chert-barite unit, but the dykes did not cut through the entire chert-barite unit, nor into the overlying pillow basalt. The top of each dyke exhibits a gradual transition into a certain chert bed, thus forming a T-junction. These relationships indicate that silica dykes might have been formed intermittently during the deposition of the chert-barite beds^{9,13,14}.

The dykes are composed mainly of fine-grained silica (less than 10 µm) with a smaller amount of sulphides and organic matter

(kerogen)⁶. The origin of the kerogen in these dykes is still controversial¹⁵, although their C and N isotope signatures and mode of occurrence imply that they may have been produced by chemotrophic organisms^{6,16}. Some silica dykes exhibit a symmetrical growth pattern along the dyke axis and occasionally have an agate, in which the silica shows fan-shape structure, grown from the hanging walls towards the centre of the dyke (Fig. 1b, c). These textures and their silica-dominated mineral assemblages indicate the precipitation of silica from a low-temperature (less than 200 °C) hydrothermal fluid^{6,13,14}.

For the analyses of fluid inclusions we used coarse-grained quartz (more than 1 mm), which shows the same growth directions as those of host silica dykes (see Supplementary Methods for sample selections). Additional samples also came from quartz veins, which are distinguished from the silica dykes by their lack of kerogen and the coarser-grained nature of the quartz. The quartz veins exhibit intrusive patterns similar to adjacent silica dykes; they would therefore have formed contemporaneously with the silica dykes.

All of our quartz samples contain H₂O–CO₂ fluid inclusions (Fig. 1e and Supplementary Fig. S2). The laser Raman spectroscopy of individual fluid inclusions revealed that these fluids consist mainly of H₂O and CO₂ with a minor but detectable amount of CH₄ (see Supplementary Notes).

The fluid inclusions are petrographically classified into types I, II and III on the basis of their distinct mode of occurrence. Type I inclusions occur as clusters along the growth faces of the host quartz, and their distribution patterns often define growth zones of the crystal (Fig. 1d and Supplementary Fig. S2a–d). Individual type I inclusions are 1–15 µm in diameter and are often oriented in the growth direction of host quartz. These textures indicate that type I represents ‘primary’ inclusion¹⁷, which was entrapped during the growth of host mineral. In contrast, type II inclusions are distributed in the plane of a secondary sealed crack, which often crosscuts more than two quartz crystals (Supplementary Fig. S2e, f). They are typically 1 µm or less in diameter. Type II inclusions are clearly entrapped after the crystallization of the host quartz and correspond to ‘secondary’ inclusions¹⁷. Type III inclusions are scattered uniformly in the host quartz without any preferred orientation. They are 5–20 µm in diameter and generally have irregular shapes (Supplementary Fig. S2g, h). Type III inclusions often occur as dense cluster in quartz and show no crack-seal texture. They are therefore probably primary inclusions, although their secondary origin cannot be completely disregarded.

The silica dykes contain type I and II inclusions, whereas the quartz veins contain type II and III. The modal fractions of secondary inclusions (type II) of all the analysed samples vary from less than 0.1 to more than 0.9 (Fig. 2a; see Supplementary Methods).

¹Research Center for the Evolving Earth and Planet, Tokyo Institute of Technology, ²Department of Earth and Planetary Science, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8551, Japan. ³Department of Environmental Science and Technology and ⁴Department of Environmental Chemistry and Engineering, Tokyo Institute of Technology, Midori-ku, Yokohama 226-8503, Japan. ⁵SORST project, Japan Science and Technology Corporation (JST), Kawaguchi, Saitama 332-0012, Japan. ⁶Department of Earth Science and Astronomy, University of Tokyo, Meguro-ku, Tokyo 153-8902, Japan.

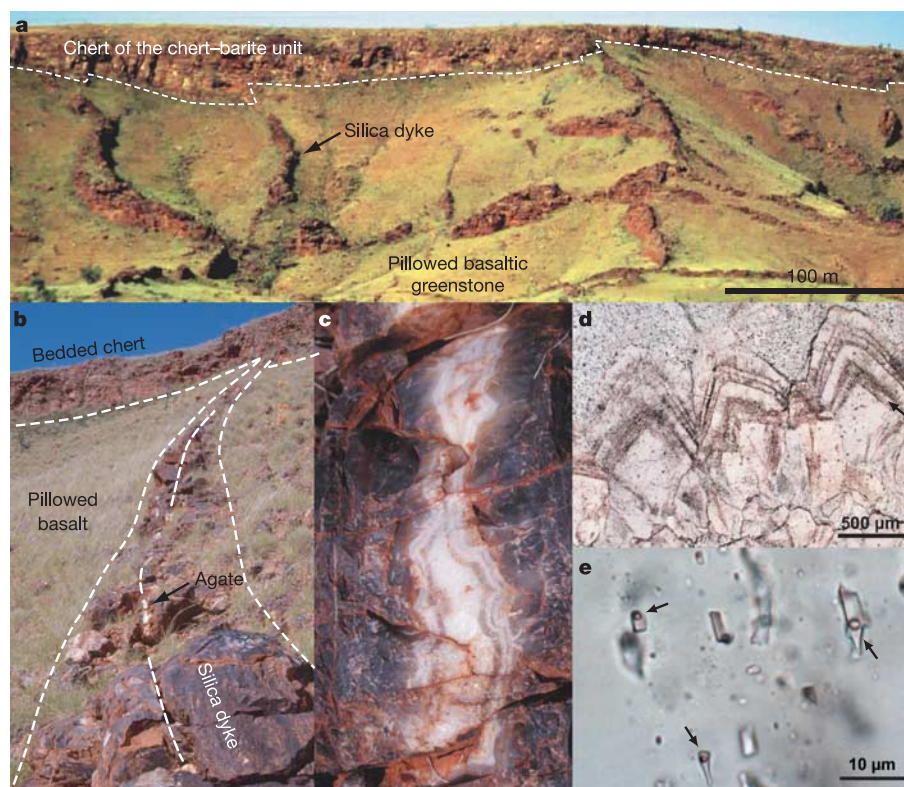


Figure 1 | Photographs of hydrothermal silica dykes and fluid inclusions therein. **a**, Annotated photograph of the Dresser Formation, showing silica dykes developed in the pillowed basaltic greenstones below the chert-barite unit (above dashed line). **b**, Photograph of an approximately 1-m-wide silica dyke. **c**, Central part of the silica dyke, showing agate and fluid-inclusion-bearing coarse-grained quartz (white portion). The black portion is

composed of fine-grained silica with organic matter. The width of the photo is about 15 cm. **d**, Optical photomicrograph of the quartz containing fluid inclusions. Numerous small black dots (arrow) are fluid inclusions distributed along the crystal faces, which define the growth zones of the host quartz. **e**, Enlarged view of the $\text{H}_2\text{O}-\text{CO}_2$ fluid inclusions (arrows). See also Supplementary Fig. S2 for the types of fluid inclusion.

The fluid inclusion volatiles were extracted from about 1 g of quartz samples by vacuum crushing (Supplementary Methods). Artificially produced CH_4 during crushing was monitored before extraction, and could be ruled out as a source of methane. The $\delta^{13}\text{C}$ values of extracted CH_4 and CO_2 are -56 to -36‰ and -7 to 0‰ , respectively (Fig. 2 and Supplementary Tables).

The $\delta^{13}\text{C}$ values of the extracted CH_4 are correlated with the modal abundance ratios of secondary inclusions (Fig. 2a). ^{13}C -depleted CH_4 is preferentially observed in samples rich in primary inclusions, whereas relatively ^{13}C -enriched CH_4 is observed in those rich in secondary inclusions. The $\delta^{13}\text{C}$ values of coexisting CO_2 also exhibit a similar systematic trend and display positive correlation with the $\delta^{13}\text{C}_{\text{CH}_4}$ values (Fig. 2c). The isotopic variations of the extracted CH_4 and CO_2 can therefore be explained by the mixing of two components: one is primary (that is, entrapped during mineral growth) and the other is secondary (that is, introduced after mineralization). The primary component has ^{13}C -depleted CH_4 and CO_2 ($\delta^{13}\text{C}_{\text{CH}_4} < -56\text{‰}$; $\delta^{13}\text{C}_{\text{CO}_2} < -4\text{‰}$), and the secondary one is more ^{13}C -enriched ($\delta^{13}\text{C}_{\text{CH}_4} \approx -35\text{‰}$; $\delta^{13}\text{C}_{\text{CO}_2} \approx 0\text{‰}$).

The origins of the CH_4 can be deduced by their isotopic compositions. Geological CH_4 has generally been categorized into three groups: microbial (emitted by the metabolic activities of methanogenic microbes), thermogenic (generated by the thermal decomposition of organic matter) and abiotic (produced by non-biological reactions from simple inorganic compounds such as CO_2 and H_2). The $\delta^{13}\text{C}_{\text{CH}_4}$ value of the secondary component ($\sim -35\text{‰}$) is within the proposed range of thermogenic CH_4 (see refs 18 and 19, for example), whereas that of the primary component (less than -56‰) is within the range of mixed gas (that is, a mixture of microbial and thermogenic CH_4 (refs 18, 19)). This empirical comparison is adequate for the first approximation; however, the origin of the

methane may be evaluated more explicitly by isotopic fractionation between the methane and potential carbon sources rather than by the $\delta^{13}\text{C}_{\text{CH}_4}$ values alone.

The CH_4 in the primary component is significantly ^{13}C -depleted relative to the coexisting CO_2 ($\Delta^{13}\text{C}_{\text{CO}_2} - \text{CH}_4 > 52\text{‰}$), which is comparable to that exhibited by microbial methanogenesis (Fig. 2b). On the basis of previous culture experiments, it can be stated that CO_2 reduction by methanogenic microbes causes distinctively large fractionations ($\Delta^{13}\text{C}_{\text{CO}_2} - \text{CH}_4 = 30\text{--}70\text{‰}$) (Fig. 2b; see ref. 20 and references therein). Although the fractionation might be smaller under high-temperature hydrothermal conditions, the large fractionation of more than 50‰ has also been observed for hyperthermophilic methanogens grown above 80°C (ref. 20). Some methanogens can also produce CH_4 by decomposing acetate, which results in fractionations of $7\text{--}22\text{‰}$ between the acetate and CH_4 (ref. 20). If it is assumed that the potential substrate acetate had a $\delta^{13}\text{C}$ value similar to that of kerogen in the silica dyke ($\sim -35\text{‰}$)⁶, the acetate fermentation process can also explain the observed fractionation of the primary component ($\Delta^{13}\text{C}_{\text{kerogen}-\text{CH}_4} > 21\text{‰}$). The methane in the primary component is therefore comparable to that produced by methanogenic microbes both by CO_2 reduction and by acetate fermentation.

Alternatively, the thermogenic origin of the primary component should also be considered because the silica dykes contain kerogen⁶. However, the lack of higher hydrocarbons (C_{2+}) in the primary component is inconsistent with its thermogenic origin (Fig. 2b). If the methane is thermogenic in origin, extracted gas should contain C_{2+} along with methane, because thermal decomposition of organic matter yields not only methane but also C_{2+} . With regard to all of the extracted fluids, ethane and propane are below the detection limit ($\text{C}_{2+}/(\text{C}_1 + \text{C}_{2+}) \ll 0.01$), which can be probably attributed to

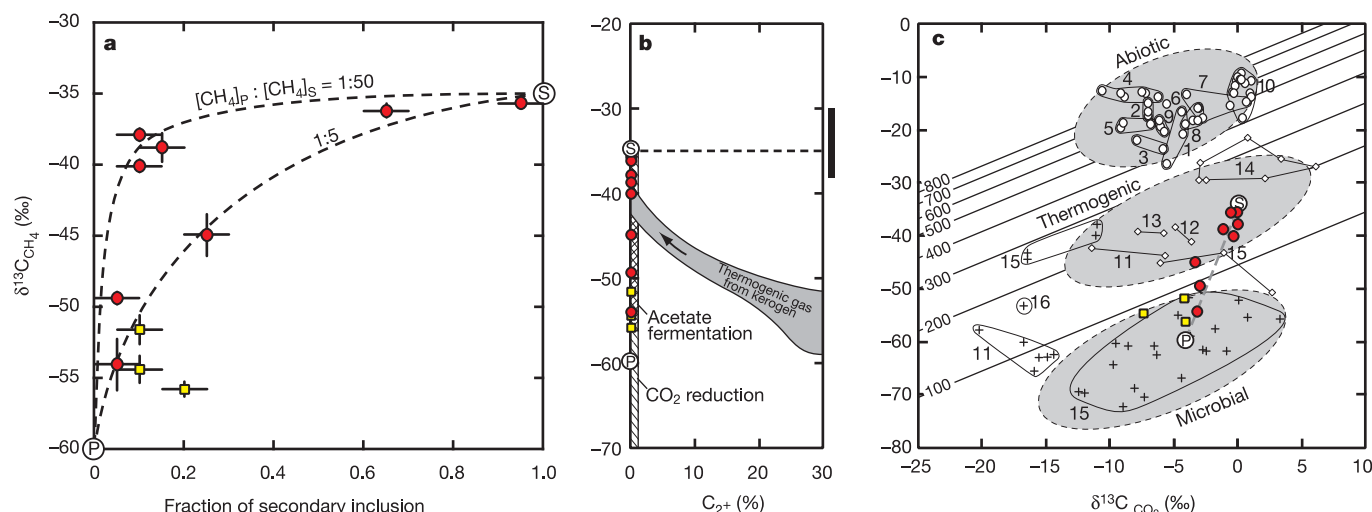


Figure 2 | Carbon isotopic compositions of CH₄ and CO₂ in the fluid inclusions. **a**, Relationship between carbon isotopic composition of extracted CH₄ and modal fraction of secondary inclusions (type II). Note that type I and III are primary inclusions occurring in silica dykes (circles) and quartz veins (squares), respectively. The two dashed lines represent mixing models between primary and secondary components, assuming that the secondary component contains a CH₄ concentration fivefold or 50-fold that of the primary component. The $\delta^{13}\text{C}_{\text{CH}_4}$ values of the primary and secondary components are assumed to be -60‰ and -35‰ , respectively. See Supplementary Note for error estimates (standard deviation of replicate runs). **b**, $\delta^{13}\text{C}$ values of the extracted CH₄ along with the expected $\delta^{13}\text{C}_{\text{CH}_4} - \% \text{C}_{2+}$ ($= 100 \times \text{C}_{2+} / (\text{C}_1 + \text{C}_{2+})$) relationship of hypothetical thermogenic gas derived from kerogen in the silica dykes. The field is calculated from a Rayleigh model²¹. The initial $\delta^{13}\text{C}$ value of kerogen is assumed to be -35‰ (dashed line), which is a representative value of

kerogen in the silica dykes (black bar⁶). The arrow indicates the directions of the isotopic and compositional changes during the maturation of the host kerogen. The oblique line and cross-hatched fields indicate the calculated ranges of bacterial CH₄ through the CO₂ reduction and acetate fermentation pathways, respectively. The calculations assume that the $\delta^{13}\text{C}$ values of substrate CO₂ and acetate are -4 and -35‰ , respectively, and fractionations are $30\text{--}70\text{‰}$ for CO₂ reduction and $7\text{--}22\text{‰}$ for acetate fermentation²⁰. **c**, Relationship between $\delta^{13}\text{C}_{\text{CH}_4}$ and $\delta^{13}\text{C}_{\text{CO}_2}$ values of the extracted fluids in comparison with those of fluids in present-day seafloor hydrothermal systems. Vent gases emitted from unsedimented ridges (circles), sedimented ridges (diamonds) and pore fluid (crosses) are interpreted to be abiotic, thermogenic and microbial, respectively. The number denotes each hydrothermal vent (see Supplementary Notes for data source). Oblique lines represent apparent temperatures in $^{\circ}\text{C}$, assuming isotopic equilibrium between CO₂ and CH₄ (ref. 27).

their microbial origin. The lack of C₂₊ might result from the decomposition of hydrocarbons into methane. However, if this is so, the $\delta^{13}\text{C}_{\text{CH}_4}$ value should be close to those of source kerogens²¹ and should therefore be much higher than -56‰ (Fig. 2b). Thermogenic origin therefore cannot explain the ^{13}C -depleted isotopic composition of the primary CH₄.

In contrast, the CH₄ in the secondary component has a more ^{13}C -enriched isotopic composition ($\sim -35\text{‰}$), which is similar to those of kerogens in silica dykes⁶ (Fig. 2b). This isotopic similarity along with the lack of C₂₊ indicates that the CH₄ might be matured thermogenic gas derived from the adjacent kerogen, although its microbial origin cannot be completely disregarded.

Last, we must consider the possibility that non-biological processes such as Fischer–Tropsch-type (FTT) reactions^{22–24} might have produced the methane. Previous experimental work has indicated that Fe–Ni-alloy-catalysed FTT reactions in hydrothermal systems can produce methane with isotopic compositions similar to those observed in this study²², but these results have not been duplicated²³. Moreover, Fe–Ni alloys and other catalysts for FTT reactions would not have existed in the surrounding rocks at temperatures prevailing during silica precipitation^{5,6}. The abiotic origin of the extracted CH₄ is therefore controversial.

Indeed, abiotic methane is emitted from various present-day seafloor hydrothermal vents^{25,26} and has a highly ^{13}C -enriched isotopic composition ($\delta^{13}\text{C}_{\text{CH}_4} = -26$ to -9‰), which is distinctively higher than those of our extracted fluids (Fig. 2c). Instead, the isotopic composition of the primary component is rather similar to those of the pore fluids in the upper part of the oceanic crust inhabited by methanogens, whereas the secondary component is similar to vent gases emitted from sedimented ridges, which are considered to be thermogenic gases (Fig. 2c). These comparisons also support the microbial origin of the primary CH₄ and the thermogenic origin of the secondary CH₄.

Thus, the fluid inclusions in the more than 3.46-Gyr-old hydrothermal silica dykes contain microbial CH₄ produced by methanogens and entrapped during the formation of the dyke. After precipitation of the host quartz, thermogenic CH₄ from thermal decomposition of the surrounding organic matter might also have been introduced into the dykes. The results of this study provide the oldest evidence of microbial methanogenesis (more than 3.46 Gyr ago), which pre-dates previous geochemical evidence³ by about 700 Myr.

Furthermore, our results demonstrate the antiquity of the Archeal domain in the universal tree of life, which is formulated by a comparison of ribosomal RNA sequences of extant organisms¹ (Supplementary Fig. S3). Methanogenesis occurs only in the Euryarchaeota, a branch of the Archaea. We therefore suggest that the domains Archaea and Bacteria would have branched from each other at least 3.46 Gyr ago. The antiquity of methanogen supports the hypothesis that microbial CH₄ would have had an important role in regulating the climate on the Archaean Earth².

Received 22 August 2005; accepted 10 January 2006.

- Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
- Kasting, J. F. & Catling, D. Evolution of a habitable planet. *Annu. Rev. Astron. Astrophys.* **41**, 429–463 (2003).
- Hayes, J. M. in *Early Life on Earth* (ed. Bengtson, S.) 220–236 (Columbia Univ. Press, New York, 1994).
- Buick, R. & Dunlop, J. S. R. Evaporitic sediments of Early Archaean age from the Warrawoona Group, North Pole, Western Australia. *Sedimentology* **37**, 247–277 (1990).
- Kitajima, K., Maruyama, S., Utsunomiya, S. & Liou, J. G. Seafloor hydrothermal alteration at Archean mid-ocean ridge. *J. Metamorph. Geol.* **19**, 583–600 (2001).
- Ueno, Y., Yoshioka, H., Maruyama, S. & Isozaki, Y. Carbon isotopes and petrography of kerogens in ~ 3.5 -Ga hydrothermal silica dikes in the North Pole area, Western Australia. *Geochim. Cosmochim. Acta* **68**, 573–589 (2004).
- Shen, Y., Buick, R. & Canfield, D. E. Isotopic evidence for microbial sulphate reduction in the early Archaean era. *Nature* **410**, 77–81 (2001).

8. Awramik, S. M., Schopf, J. W. & Walter, M. R. Filamentous fossil bacteria from the Archean of Western Australia. *Precamb. Res.* **20**, 357–374 (1983).
9. Ueno, Y., Isozaki, Y., Yurimoto, H. & Maruyama, S. Carbon isotopic signatures of individual Archean microfossils (?) from Western Australia. *Int. Geol. Rev.* **43**, 196–212 (2001).
10. Buick, R. Microfossil recognition in Archean rocks: an appraisal of spheroids and filaments from a 3500 M.Y. old chert-barite unit at North Pole, Western Australia. *Palaios* **5**, 441–491 (1990).
11. Garcia Ruiz, J. M. et al. Self-assembled silica-carbonate structures and detection of ancient microfossils. *Science* **302**, 1194–1197 (2003).
12. Thorpe, R. I., Hickman, A. H., Davis, D. W., Mortensen, J. K. & Trendall, A. F. in *The Archean: Terrains, Processes and Metallogeny* (eds Glover, J. E. & Ho, S. E.) 395–406 (Univ. Western Australia Publ. 9, Perth, 1992).
13. Nijman, W., de Bruijne, K. H. & Valkering, M. E. Growth fault control of Early Archean cherts, barite mounds and chert-barite veins, North Pole Dome, Eastern Pilbara, Western Australia. *Precamb. Res.* **95**, 247–274 (1999).
14. Van Kranendonk, M. J. & Pirajno, F. Geochemistry of metabasalts and hydrothermal alteration zones associated with c. 3.45 Ga chert and barite deposits: implications for the geological setting of the Warrawoona Group, Pilbara Craton, Australia. *Geochem. Explor. Envir. Anal.* **4**, 253–278 (2004).
15. Brasier, M. D. et al. Questioning the evidence for Earth's oldest fossils. *Nature* **416**, 76–81 (2002).
16. Pinti, D. L., Hashizume, K. & Matsuda, J. Nitrogen and argon signatures in 3.8 to 2.8 Ga metasediments: clues on the chemical state of the Archean ocean and the deep biosphere. *Geochim. Cosmochim. Acta* **65**, 2301–2315 (2001).
17. Roedder, E. *Fluid Inclusions* (Reviews in Mineralogy vol. 12, Mineralogical Society of America, Washington DC, 1984).
18. Schoell, M. Genetic characterization of natural gas. *Am. Assoc. Petrol. Geol. Bull.* **67**, 2225–2238 (1983).
19. Whiticar, M. J. Carbon and hydrogen isotope systematics of bacterial formation and oxidation of methane. *Chem. Geol.* **161**, 291–314 (1999).
20. Valentine, D. L., Chidthaisong, A., Rice, A., Reeburgh, W. S. & Tyler, S. C. Carbon and hydrogen isotope fractionation by moderately thermophilic methanogens. *Geochim. Cosmochim. Acta* **68**, 1571–1590 (2004).
21. Clayton, C. Carbon isotope fractionation during natural gas generation from kerogen. *Mar. Petrol. Geol.* **8**, 232–240 (1991).
22. Horita, J. & Berndt, M. E. Abiogenic methane formation and isotopic fractionation under hydrothermal conditions. *Science* **285**, 1055–1057 (1999).
23. McCollom, T. M. & Seewald, J. S. A reassessment of the potential for reduction of dissolved CO₂ to hydrocarbons during serpentinization of olivine. *Geochim. Cosmochim. Acta* **65**, 3769–3778 (2001).
24. Sherwood Lollar, B., Westgate, T. D., Ward, J. A., Slater, G. F. & Lacrampe-Couloume, G. Abiogenic formation of alkanes in the Earth's crust as a minor source for global hydrocarbon reservoirs. *Nature* **416**, 522–524 (2002).
25. Welhan, J. A. Origin of methane in hydrothermal systems. *Chem. Geol.* **71**, 183–198 (1988).
26. Charlou, J. L., Donval, J. P., Fouquet, Y., Jean-Baptiste, P. & Holm, N. G. Geochemistry of high H₂ and CH₄ vent fluids issuing from ultramafic rocks at the Rainbow hydrothermal field (36°14'N, MAR). *Chem. Geol.* **191**, 345–359 (2002).
27. Richet, P., Bottinga, Y. & Javoy, M. A review of hydrogen, carbon, nitrogen, oxygen, sulphur, and chlorine stable isotope fractionation among gaseous molecules. *Annu. Rev. Earth Planet. Sci.* **5**, 65–110 (1977).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Terabayashi, Y. Kato, K. Okamoto, T. Ota, T. Kabashima, K. Kitajima and K. Shimizu for assistance in field work, A. Thorne, K. J. McNamara and A. H. Hickman for field collaboration, H. Nara, Y. Matsui and M. Nishizawa for assisting in the construction of the vacuum line, and R. Buick and J. F. Kasting for comments on early versions of this manuscript. This research was supported by the 21st Century COE Program 'How to build habitable planets,' Tokyo Institute of Technology, sponsored by the Ministry of Education, Culture, Sports, Technology and Science, Japan. Y.U. thanks the Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.U. (yueno@depe.titech.ac.jp).

LETTERS

Modelling conservation in the Amazon basin

Brítaldo Silveira Soares-Filho^{1,2}, Daniel Curtis Nepstad^{3,5,6}, Lisa M. Curran⁴, Gustavo Coutinho Cerqueira¹, Ricardo Alexandrino Garcia², Claudia Azevedo Ramos⁵, Eliane Voll¹, Alice McDonald⁴, Paul Lefebvre³ & Peter Schlesinger³

Expansion of the cattle and soy industries in the Amazon basin has increased deforestation rates and will soon push all-weather highways into the region's core¹⁻⁴. In the face of this growing pressure, a comprehensive conservation strategy for the Amazon basin should protect its watersheds, the full range of species and ecosystem diversity, and the stability of regional climates. Here we report that protected areas in the Amazon basin—the central feature of prevailing conservation approaches⁵⁻⁸—are an important but insufficient component of this strategy, based on policy-sensitive simulations of future deforestation. By 2050, current trends in agricultural expansion will eliminate a total of 40% of Amazon forests, including at least two-thirds of the forest cover of six major watersheds and 12 ecoregions, releasing 32 ± 8 Pg of carbon to the atmosphere. One-quarter of the 382 mammalian species examined will lose more than 40% of the forest within their Amazon ranges. Although an expanded and enforced network of protected areas could avoid as much as one-third of this projected forest loss, conservation on private lands is also essential. Expanding market pressures for sound land management and prevention of forest clearing on lands unsuitable for agriculture are critical ingredients of a strategy for comprehensive conservation^{3,4}.

The Amazon has entered a new era as the growing profitability of cattle ranching and soy production increases deforestation rates and drives the expansion of the highway network into the region's core¹⁻⁴ (Supplementary Fig. S1). The ecological effects of this new phase of accelerated deforestation are potentially large. For example, Amazon trees contain 119 ± 28 Pg of carbon⁹, equivalent to 1.5 decades of current worldwide anthropogenic carbon emissions to the atmosphere. Both regional and global climate systems are also coupled to the Amazon forest through latent heat transfer^{10,11}.

Conservation strategies for the Amazon region have focused on protected areas (PAs), both inhabited ('extractive reserves', public forests and indigenous lands) and uninhabited (parks and biological reserves)⁵⁻⁸. Regional workshops⁸ and gap assessments¹² have developed conservation priorities and determined the representation of vegetation types and centres of endemism within current and proposed PAs. PAs inhibit both deforestation and fire¹³ but they are less effective in conserving watersheds, whose headwaters generally extend beyond reserve boundaries. Watershed protection depends on the maintenance of riparian zone vegetation, required by law on private properties in Brazil. Nor will PAs be sufficient to sustain Amazon climate; more than 70% of the forest cover of Amazon landscapes may be necessary to maintain the forest-dependent rainfall regime¹¹.

We compared the potential influence of PAs and other conservation approaches on future trends in Amazon watersheds, vegetation types (ecoregions), mammals and carbon emissions by developing an

empirically based, policy-sensitive model of Amazon deforestation. The model was run under eight scenarios that encompass a plausible range of future trajectories of deforestation. At one extreme is the 'business-as-usual' scenario (BAU), which assumes that: recent deforestation trends will continue; highways currently scheduled for paving will be paved; compliance with legislation requiring forest reserves on private land will remain low; and new PAs will not be created. The BAU scenario assumes that as much as 40% of the forests inside of PAs are subject to deforestation (B.S.S.-F., unpublished observation), climbing to 85% outside.

At the other extreme, the 'governance' scenario assumes that Brazilian environmental legislation is implemented across the Amazon basin through the refinement and multiplication of current experiments in frontier governance^{3,4}. These experiments include enforcement of mandatory forest reserves on private properties through a satellite-based licensing system¹⁴, agro-ecological zoning of land use¹⁵, and the expansion of the PA network (Amazon Region Protected Areas Program¹⁶). The plausibility of a scenario of expanding frontier governance is demonstrated by growing pressures on Amazon cattle ranchers and soy farmers from international markets and financial institutions to comply with environmental legislation and manage their land soundly⁴, recent successes in designating new PAs in regions of active frontier expansion (for example, 7 million ha of new reserves were created in active frontier regions in 2004 and 2005) and the regional, participatory planning processes that are preceding the paving of new highways between Santarém and Mato Grosso¹⁷, and in the southwestern Amazon¹⁸. Within the governance scenario, the deforestation rate, although rising initially owing to road paving, declines over time, simulating the effects of growing market pressures in favour of sound land management, emerging markets for carbon retained in native forests¹⁹ and other incentives for landholders who conserve forest on their properties⁴. The governance scenario assumes that the planned expansion of the PA network in the Brazilian Amazon¹⁶, from 32% to 41% of the total forest area, succeeds and 100% of the forests in PAs are preserved intact, with only 50% of the forests outside of PAs subject to deforestation (compared with 20% currently permitted by Brazilian regulations). We disaggregated the effects of highway paving by running the model with no further paving and with three additional scenarios in which a single highway is paved: the BR-163 (Cuiabá–Santarém) and the Interoceanica (Assis Brasil–Cuzco) paved in 2008 and the Manaus–Porto Velho highway paved in 2010. These scenarios are summarized in Supplementary Table S3. The scenarios are conservative because they do not consider forest impoverishment through logging and fire²⁰, nor do they consider the potential for forest substitution by savanna scrub through global warming²¹. Moreover, we did not include the loss of Amazon savannas.

¹Centro de Sensoriamento Remoto, and ²Centro de Desenvolvimento e Planejamento Regional, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Belo Horizonte, 31270-901, MG, Brazil. ³The Woods Hole Research Center, Woods Hole, Massachusetts 02543-0296, USA. ⁴Tropical Resources Institute, Yale School of Forestry and Environmental Studies, 370 Prospect Street, New Haven, Connecticut 06511, USA. ⁵Instituto de Pesquisa Ambiental da Amazônia, Av. Nazaré 669, 66035-170, Belém, Pará, Brazil. ⁶Universidade Federal do Pará, Núcleo de Altos Estudos Amazônicos, Campus do Guamá, 66075-110, Belém, Pará, Brazil.

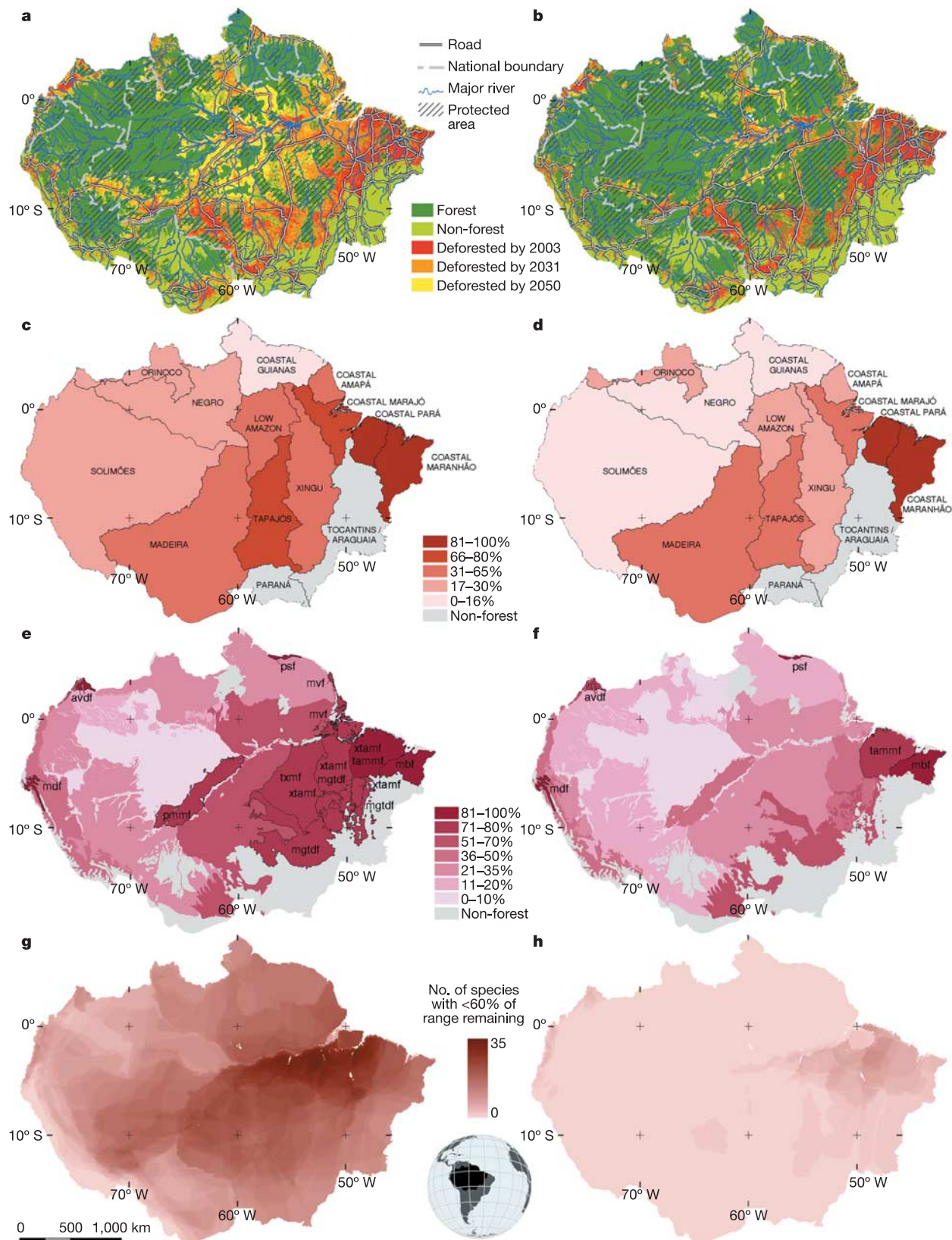


Figure 1 | Model results for the extreme-case scenarios in the year 2050. **a, b**, Forest cover for BAU (**a**) and governance (**b**) scenarios. **c, d**, Percentage forest loss by major watershed for BAU (**c**) and governance (**d**) scenarios in 2050. **e, f**, Percentage forest loss by ecoregion for BAU (**e**) and governance (**f**) scenarios in 2050. **g, h**, Numbers of imperilled mammals for BAU (**g**) and governance (**h**) scenarios in 2050 ($n = 105$). Ecoregions: avdf,

Apure/Villavicencio dry forests; mbf, Maranhão Babaçu forests; mdf, Marañon dry forests; mgtdf, Mato Grosso tropical dry forests; mvf, Marajo Varzea forests; nr, Northeastern restingas; pm, Pará mangroves; pmmf, Purus/Madeira moist forests; psf, Paramaribo swamp forests; tammf, Tocantins-Araguaia/Maranhão moist forests; txmf, Tapajos/Xingu moist forests; xtamf, Xingu/Tocantins-Araguaia moist forests.

We estimate that the closed-canopy forest formation of the Amazon will be reduced from its current area of 5.3 million km² (2003, 85% of the original area) to 3.2 million km² (53%) by 2050 if current trends continue unabated (BAU) (Fig. 1a and Supplementary Fig. S7). Under the governance scenario, 4.5 million km² of forest would remain in 2050 (Fig. 1b and Supplementary Fig. S7). The intermediate scenarios that we ran in the model indicate that simply expanding the PA network, but with lax enforcement, reduces new deforestation 7% below the BAU scenario baseline. All conservation measures combined (but without PA expansion) accounted for 86% of the deforestation that is avoided in the 'governance' scenario. An expanded network of PAs that is effectively implemented and enforced provided half of this reduction in deforestation.

Paving of the Manaus–Porto Velho highway, which traverses a region with few PAs and little human settlement, would stimulate more deforestation than the paving of either the Cuiabá–Santarém or the Interoceanica highways alone (Supplementary Figs S11–S13 and Supplementary Tables S5–S7). The paving of the BR-163 highway would result in an extensive region of deforestation as the agricultural frontier moving northwest from São Felix do Xingu coalesces with the eastward expansion from the BR-163. Paving of the Interoceanica highway would cause the least impact, but would provoke extensive deforestation in the southwestern Brazilian state of Acre.

In all scenarios, future deforestation is concentrated in the eastern Amazon, where the density of paved highways will continue to be highest for several decades, and along the BR-364 highway from Rondônia into Acre in the Southwestern Amazon (Fig. 1a, b). With the exception of Santa Cruz (Bolivia), Pucallpa (Peru), the Florencia region of Colombia, and coastal regions of the Guianas, large blocks of forest outside Brazil and most of the northwest quadrant of the Brazilian Amazon may remain largely intact until the mid-century, 'passively' protected by their remoteness. However, our analysis considers only existing or planned highways. Our projections of future deforestation would increase if additional highways were constructed into currently inaccessible regions.

Given the concentration of projected deforestation in the eastern Amazon, some watersheds, ecoregions and mammalian species are far more vulnerable to disruption than others. Eight of the 12 major watersheds of the Amazon will lose more than half of their forest cover under the 2050 BAU scenario. Some Atlantic coastal watersheds (Pará, Maranhão, Marajó and Amapá) and southeastern tributaries of the Amazon River (Tapajós and Xingu) will all lose at least two-thirds of their forest cover (Fig. 1c, d, and Supplementary Table S8) and may undergo substantial increases in peak river flow and flooding as pastures replace forests^{22,23}. Eighteen of the Amazon's 32 major forested ecoregions will lose more than 40% of their forest cover by 2050 and 12 will lose more than 70% (Fig. 1e, f, and Supplementary Table S9). The most vulnerable ecoregions are found between savanna woodlands and closed-canopy forest where human occupation is concentrated and PAs are scarce. These include the Maranhão babaçu forest (97% loss) and the 420,000 km² Mato Grosso dry forest (76% loss). Wetland forests are also highly vulnerable, such as the Paramaribo swamp forest (93% forest loss) and the Marajó forest (78%). These ecoregions are highly dependent on the successful protection and possible expansion of a few indigenous reserves (for example, *Parque Indígena do Xingu*) and parks (for example, *Parque Nacional do Gurupi*)¹³.

Terrestrial non-flying mammals serve as conservative indicator taxa or proxies to estimate how future deforestation will affect individual species. The highest concentration of 'imperilled' mammals—those species that lose more than 40% of the forest in their Amazon ranges—is in the east-central Amazon (Fig. 1g, h). Thirty-five primate species lose 60–100% of their Amazonian range. Hence, although the highest levels of mammal endemism and species richness are found in the southwestern Amazon close to the Andes²⁴ (Supplementary Fig. S14), the larger mammalian extinction threat²⁵

may occur in the eastern Amazon, where projected deforestation rates are much higher.

Carbon emissions expected for each scenario were estimated by assuming that 85% of the carbon contained in forest trees is released to the atmosphere after deforestation²⁶ and by superimposing deforestation simulations on a range of forest biomass maps⁹, extrapolated to the entire Amazon. Uncertainty estimates are described in Supplementary Information. By 2050, 32 ± 8 Pg of carbon is emitted under the BAU scenario, equivalent to four years of current annual emissions worldwide, contrasted with 15 ± 4 Pg under the governance scenario.

The conservation measures simulated in the 'governance' scenario would reduce the number of imperilled watersheds, ecoregions and mammalian species by about two-thirds and would avoid carbon emissions to the atmosphere equivalent to two years of global human-induced emissions. Conservation achievements of this magnitude are unlikely to result from command-and-control implementation of environmental legislation alone, but instead become more likely as international markets impose higher environmental standards on beef, soy and other food commodities. Compliance with Brazil's environmental legislation requiring the protection of riparian vegetation and forest reserves on private land could increase as cattle ranchers and soy farmers perceive sound land management to be a requirement to access lucrative international markets⁴. The economic opportunity costs of reducing deforestation in the Amazon could be lowered by restricting forest clearing on lands with low potential for crop production or cattle ranching, which are abundant in the region³.

Current legislation permits landholders to exceed the limit on forest clearing on their property if they retain other properties on which forest reserves comprise more than this limit. Land-use zoning systems under development by Amazon states could reinforce a system of deforestation licensing that restricts the clearing of lands with low productive potential as it relaxes restrictions on lands with high productive potential.

With many of the benefits of Amazon conservation accruing to humanity worldwide, developed countries must be willing to pay to make frontier governance politically feasible. Beyond the market incentives for ecologically sound land management that may be captured through the environmental certification of beef, soybeans and timber⁴, part of this funding might be provided through the sale of carbon credits derived from the avoidance of 17 ± 4 Pg of carbon emissions within a modified climate change convention¹⁹. The potential decrease in Amazon carbon emissions within the governance scenario is more than eightfold the worldwide decrease in greenhouse gas emissions that will be achieved during the first compensation period of the Kyoto protocol. Recent advances in enforcing the region's ambitious deforestation legislation¹⁴ and regional planning exercises underway along the major highways planned for paving^{17,18} are just two of the large-scale conservation efforts that could be the target of investments made for lowering the region's carbon emissions.

The Amazon PA network may protect a large portion of the region's mammalian diversity, but will not avoid the impoverishment of critical watersheds and ecoregions. An expanded conservation strategy must encompass lands that fall outside PAs if it is to avoid the collapse of regional rainforest ecosystems that is already occurring elsewhere in the tropics²⁷.

METHODS

The model, which we call 'SimAmazonia 1' (from the Portuguese *simulação da Amazônia*) produces annual maps of simulated future deforestation under user-defined scenarios of highway paving, PA networks, PA effectiveness, deforestation rates and deforested land ceilings. We stratified the Amazon basin into 47 socioeconomic subregions for which individualized deforestation rates are forecast; these rates were estimated from historical trends (from 1997 to 2002) derived from satellite-based deforestation maps (Supplementary Table S2), the

planned road paving schedule (Supplementary Table S1) and existing and proposed PAs (Supplementary Table S2). Proximity to paved highways is the major driver of deforestation rates in the model and this relationship was defined empirically from data on deforestation and paved roads for 432 counties of the Brazilian Amazon (Supplementary Fig. S5). Increasing proximity to paved highways accelerates deforestation within a subregion up to an inflection point when forests outside protected areas start becoming scarce. The spatial distribution of deforestation across the Amazon is simulated with a cellular automata model, with parameters customized for each subregion, that allocates deforestation on the basis of its empirical relationships with proximity to roads, rivers and towns, land use zoning, and biophysical features²⁸ represented in raster grids of 1 km² resolution. Spatial integrity across subregions is attained by employing spatial variables (for example, distance to previously deforested land and distance to all roads) that are updated annually over the entire basin. This latter variable is output of the road constructor model²⁸, a component that simulates the expansion of the secondary road network and thereby incorporates the effect of endogenous²⁹ roads on the evolving spatial patterns of deforestation. The spatial simulation model was calibrated and validated for 12 regional case studies (B.S.S.-F., unpublished material), each represented by a Landsat Thematic Mapper scene (180 km × 180 km). A detailed description of the model design is provided in Supplementary Information.

Received 21 September; accepted 4 November 2005.

- Kaimowitz, D., Mertens, B., Wunder, S. & Pacheco, P. *Hamburger Connection Fuels Amazon Destruction* [online] (http://www.cifor.cgiar.org/publications/pdf_files/media/Amazon.pdf) (2004).
- Alencar, A. et al. *Desmatamento na Amazônia: Indo Além da 'Emergência Crônica'* [online] (http://www.ipam.org.br/publicacoes/livros/resumo_desmatamento.php) (2004).
- Nepstad, D. et al. Frontier governance in Amazonia. *Science* **295**, 629–631 (2002).
- Nepstad, D. et al. The economic “teleconnections” of the Amazon beef and soy industries: opportunities for conservation. *Conserv. Biol.* (in the press).
- Peres, C. A. & Terborgh, J. W. Amazonian nature reserves: an analysis of the defensibility status of existing conservation units and design criteria for the future. *Conserv. Biol.* **9**, 34–46 (1995).
- Verissimo, A., Cochrane, M. A. & Souza, C. Jr. National forests in the Amazon. *Science* **297**, 1478 (2002).
- Schwartzman, S., Moreira, A. G. & Nepstad, D. C. Rethinking tropical forest conservation: perils in parks. *Conserv. Biol.* **14**, 1351–1357 (2000).
- Capobianco, J. P. R. et al. *Biodiversidade na Amazônia Brasileira: Avaliação e Ações Prioritárias para a Conservação, Uso Sustentável e Repartição dos Benefícios* (Instituto Socioambiental, São Paulo, Brazil, 2001).
- Houghton, R. A., Lawrence, K. T., Hackler, J. & Brown, L. S. The spatial distribution of forest biomass in the Brazilian Amazon: a comparison of estimates. *Glob. Change Biol.* **7**, 731–746 (2001).
- Werth, D. & Avissar, R. The local and global effects of Amazon deforestation. *J. Geophys. Res.* **107**, 8087 (doi:10.1029/2001JD000717) (2002).
- Silva Dias, M. A. F. et al. Cloud and rain processes in biosphere-atmosphere interaction context in the Amazon region. *J. Geophys. Res.* **107**, 8072 (doi:10.1029/2001JD00335) (2002).
- Fearnside, P. M. & Ferraz, J. A. Conservation gap analysis of Brazil's Amazonian vegetation. *Conserv. Biol.* **9**, 1134–1147 (1995).
- Nepstad, D. et al. Inhibition of Amazon deforestation and fire by parks and indigenous reserves. *Conserv. Biol.* (in the press).
- Fearnside, P. Deforestation control in Mato Grosso: a new model for slowing the loss of Brazil's Amazon forest. *Ambio* **32**, 343–345 (2003).
- IMAC. *Zoneamento Ecológico-Econômico do Acre* [online] (http://www.ac.gov.br/m_amb/zee.htm) (2000).
- Montiel, F. *Programa Áreas Protegidas da Amazônia—ARPA. Resumo Executivo* [online] (<http://www.mma.gov.br/port/sca/ppg7/doc/arpaeex.pdf>) (2004).
- Alencar, A. et al. *O Desenvolvimento que Queremos: Ordenamento Territorial da Br-163, Baixo Amazonas, Transamazônica e Xingu* (IPAM, ISA, FVPP, FETAGRI-BAM, FORMAD, FVPP, CEFTBAM, Fórum dos Movimentos Sociais da BR-163, GTA, Santarém, Brazil, 2004).
- MAP. *The Madre de Dios, Acre, and Pando Initiative* [online] (<http://map-amazonia.net/>) (2005).
- Santilli, M. P. et al. Tropical deforestation and the Kyoto Protocol: an editorial essay. *Clim. Change* **71**, 267–276 (2005).
- Nepstad, D. C. et al. Large-scale impoverishment of Amazonian forests by logging and fire. *Nature* **398**, 505–508 (1999).
- Oyama, M. D. & Nobre, C. A. A new climate–vegetation equilibrium state for Tropical South America. *Geophys. Res. Lett.* **30**, 2199 (doi:10.1029/2003GL018600) (2003).
- Nobre, C. A. et al. in *Amazonian Deforestation and Climate* (eds Gash, J. H. C., Nobre, C. A., Roberts, J. R. & Victoria, R. L.) 577–596 (Wiley, Chichester, 1996).
- Costa, M. H., Botta, A. & Cardille, J. A. Effects of large-scale changes in land cover on the discharge of the Tocantins River, Southeastern Amazonia. *J. Hydrol.* **283**, 206–217 (2003).
- Mittermeier, R., Myers, N., Mittermeier, C. G., Fonseca, G. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
- Tilman, D., May, R. M., Lehman, C. L. & Nowak, M. A. Habitat destruction and the extinction debt. *Nature* **371**, 65–66 (1994).
- Houghton, R. A. et al. Annual fluxes of carbon from deforestation and regrowth in the Brazilian Amazon. *Nature* **403**, 301–304 (2000).
- Curran, L. M. et al. Lowland forest loss in protected areas of Indonesian Borneo. *Science* **303**, 1000–1003 (2004).
- Soares-Filho, B. S. et al. Simulating the response of land-cover changes to road paving and governance along a major Amazon highway: The Santarém–Cuiabá corridor. *Glob. Change Biol.* **10**, 745–764 (2004).
- Souza, C. Jr, Brandão, A. Jr, Anderson, A. & Verissimo, A. *Avanço das Estradas Endógenas na Amazônia* [online] (<http://www.imazon.org.br/upload/124466.pdf>) (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Schwartzmann for comments on the manuscript. This work was supported by grants from the Gordon and Betty Moore Foundation, the National Aeronautics and Space Administration (Large-Scale Biosphere Atmosphere Experiment), NASA Earth Science Program, the National Science Foundation, the Santa Fe Institute, and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

Author Contributions B.S.S.-F. and D.C.N. designed the model and did the carbon analysis. G.C.C. programmed the model software. R.A.G. performed the socioeconomic analyses. L.M.C., C.A.R. and A.M. analysed the ecoregions and mammal geographic distributions. E.V., P.L. and P.S. prepared the database.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.S.S.-F. (britaldo@csr.ufmg.br) or D.C.N. (dneptad@whrc.org).

LETTERS

The nature of plant species

Loren H. Rieseberg¹, Troy E. Wood¹ & Eric J. Baack¹

Many botanists doubt the existence of plant species^{1–5}, viewing them as arbitrary constructs of the human mind, as opposed to discrete, objective entities that represent reproductively independent lineages or ‘units of evolution’. However, the discreteness of plant species and their correspondence with reproductive communities have not been tested quantitatively, allowing zoologists to argue that botanists have been overly influenced by a few ‘botanical horror stories’, such as dandelions, blackberries and oaks^{6,7}. Here we analyse phenetic and/or crossing relationships in over 400 genera of plants and animals. We show that although discrete phenotypic clusters exist in most genera (>80%), the correspondence of taxonomic species to these clusters is poor (<60%) and no different between plants and animals. Lack of congruence is caused by polyploidy, asexual reproduction and over-differentiation by taxonomists, but not by contemporary hybridization. Nonetheless, crossability data indicate that 70% of taxonomic species and 75% of phenotypic clusters in plants correspond to reproductively independent lineages (as measured by postmating isolation), and thus represent biologically real entities. Contrary to conventional wisdom⁸, plant species are more likely than animal species to represent reproductively independent lineages.

Previous attempts to assess the discreteness or objectivity of plant species have relied on evidence from floras⁹, monographs¹⁰ and the degree of correspondence between folk and western taxonomies^{11,12}. However, these studies are inconclusive because of potential biases associated with human neurological processes^{13,14}. Also, they fail to address whether discrete clusters, if they exist, correspond to reproductively independent lineages.

A more rigorous approach derives from the application of statistical procedures to classification—so-called numerical taxonomy¹⁵. To bring objectivity into taxonomic practices, numerical taxonomists advocated a classification based on the quantitative analyses of as many characters as possible, with each character given equal weighting. Although now largely replaced by phylogenetic methods, these statistical practices continue to be broadly employed by taxonomists working on species-level problems.

We surveyed the biosystematic literature for numerical taxonomic studies of plants and animals that sampled multiple populations per species taxon (see the Methods). For each of 218 such studies that were identified (Supplementary Tables 1 and 2), we tabulated the number of discrete phenotypic clusters as revealed by statistical methods. We then calculated the proportion of taxonomic species (based on the last taxonomic treatment of that group before phenetic analyses) that correspond directly to discrete phenotypic clusters. To identify biological factors that influence species discreteness, multi-factor analysis of variance (ANOVA) was used to test for significant effects of taxon (division, class), life history, mating system, polyploidy and contemporary hybridization on the proportion of species that correspond directly to phenetic clusters (see the Methods).

In the majority of genera for which five or more species taxa were assessed, phenetic analyses revealed discrete clusters: 83% for plants

($n = 30$) and 88% for animals ($n = 9$). Thus, phenotypic discontinuities do exist in most taxonomic groups. However, the percentage of species taxa that correspond directly to these clusters was low (52.8% for plants and 52.1% for animals), particularly when compared to earlier estimates of the proportion of ‘good’ plant species from floristic (83.1%)⁹ and monographic (93%)¹⁰ studies. Lack of correspondence was mostly due to over-differentiation (>1 species taxon per cluster) by taxonomists (87.4%) rather than under-differentiation (13.6%).

So how do we account for the poor concordance between species taxa and phenotypic clusters? Possibly, taxonomists are too willing to split taxa on the basis of one or a few traits that ostensibly discriminate groups. Lack of correspondence may also have a biological basis, a hypothesis supported by ANOVA conducted on the plant and animal phenetic data (Table 1; Fig. 1). While taxonomic group, life history and contemporary hybridization had no effect on correspondence between species taxa and phenetic clusters, asexual reproduction and polyploidy reduced correspondence. Because both asexuality and polyploidy are attended by significant crossing barriers, phenotypically intermediate hybrids in these groups may persist without recombining with parental forms¹⁶. However, the importance of these factors in botanical classification may be exaggerated in our study by a bias towards problematic taxa in the application of phenetic methods; our compilation included 21 phenetic studies of agamic complexes (13.1%) compared to an expectation of ~1% if taxonomic effort were distributed randomly across genera¹⁷. More generally, groups that do not contain polyploid or agamic taxa, yet have been a source for taxonomic confusion, are also heavily represented in our data set. However, there was not a significant effect of these difficult taxa on correspondence (model $F_{1,164} = 1.20$; $P = 0.28$; see the Supplementary Methods and Results).

A more surprising observation was that contemporary hybridization among species of the same ploidal level failed to cause taxonomic problems, despite its frequent mention as the primary cause of ‘fuzzy’ species-boundaries in plants⁸. Perhaps diploid hybrids rarely cause taxonomic problems because they tend to be pulled back into the orbit of the parental species by backcrossing. Indeed, another study also reported that hybridization was rarely associated with

Table 1 | ANOVA of factors affecting correspondence between species taxa and phenetic clusters

Source	d.f.*	Sum of squares	F	P
Taxon	3	0.38	0.81	0.49
Life history	1	0.04	0.28	0.60
Mating system	2	1.63	5.20	0.0065
Polyploidy	1	1.16	7.39	0.0073
Hybridization	1	0.32	2.07	0.15
Error	159	25.02	–	–
Asexuality versus sexuality	1	1.64	10.40	0.0015
Selfing versus outcrossing	1	0.51	3.24	0.074

*d.f., degrees of freedom.

¹Department of Biology, Indiana University, Bloomington, Indiana 47405, USA.

problematic taxa¹⁰. Finally, we found marginal statistical support (see Table 1) for Baker's¹⁶ hypothesis that selfing lineages are more likely to be separated by phenotypic discontinuities than are outcrossing taxa.

To determine whether species taxa and/or phenetic clusters also represent reproductively independent lineages, as measured by postmating isolation, we searched the plant biosystematic literature for crossing studies involving the same genera employed in the phenetic analyses. If hybrids from intraspecific/intracluster crosses were fertile and viable, but crosses with closely related congeners were significantly less successful, then the species/cluster probably represents a reproductively independent lineage. Thus, we calculated a crossability index (CI)¹⁸ for each interspecific (or intercluster) cross-combination within a given study, by dividing the mean interspecific (or intercluster) crossability by the average intraspecific (or intracluster) crossability. Interspecific crossability was considered to be significantly lower than intraspecific crossability if CIs were reduced by two standard deviations of intraspecific crossability differences within a given group, typically a CI of ~ 0.8 . Counting a species taxon (or phenetic cluster) as a reproductively independent lineage required that CIs from all cross-combinations involving that species or cluster fall below this threshold.

Analyses of 37 taxonomic groups having both phenetic and crossability data (Supplementary Table 3) revealed that both species taxa ($71.2 \pm 7.1\%$; mean $\pm$ s.e.m.) and phenetic clusters ($75.2 \pm 6.8\%$) generally exhibit reproductive independence, and a paired *t*-test failed to detect a difference in means ($n = 37$, $t = 0.12$, $P = 0.91$; two-tailed paired *t*-test). However, there was a trend for taxonomists to outperform phenetic clustering when polyploids were present ($n = 9$; $t = 1.44$, $P = 0.08$; one-tailed paired *t*-test). This latter result is expected because phenetic clustering fails to take into account variation in ploidy, which often generates strong reproductive isolation. In addition, single factor regression, weighted by the number of species per genus, indicated that mean CI had a strong influence on the fraction of species taxa that correspond to phenetic clusters (model $F_{1,28} = 7.93$, $P = 0.009$; $R^2 = 0.22$). That is, taxa that are strongly reproductively isolated also tend to be distinct phenotypically, an observation which is consistent with an important

role for reproductive barriers in the formation and maintenance of discrete morphological groups.

While these results imply the majority of plant species represent reproductively independent lineages, the number of taxa included in this initial analysis is small. Therefore, we extended the crossability analysis (see the Methods) to a much broader array of plant and animal groups (in the absence of phenetic data, we assume species represent morphologically distinct units), including 114 plant genera representing 1,231 interspecific cross-combinations and 170 genera of animals representing 694 interspecific cross-combinations (Supplementary Tables 4–6).

Contrary to expectations⁸, plant species taxa were significantly more likely to represent reproductively independent lineages (as measured by postmating isolation) than animals ($69.5 \pm 3.7\%$ versus $39.2 \pm 4.3\%$; model $F_{1,306} = 42.30$, $P < 0.0001$). However, there was considerable heterogeneity among plant and animal groups in the fraction of species taxa that exhibit reproductive independence (model $F_{7,299} = 6.61$, $P < 0.0001$), with ferns and fern allies showing the highest levels of independence, and birds the lowest (Fig. 2). Surprisingly, none of the biological factors considered in the phenetic analyses (see the Methods) had a significant impact on the fraction of species that represent reproductively independent lineages.

Analysis of variation in the strength of postmating reproductive isolation also revealed that plant species taxa were, on average, more strongly isolated than animal species (mean CI for plants 0.43 ± 0.03 , versus 0.71 ± 0.03 for animals; model $F_{1,278} = 39.97$, $P < 0.0001$). This result should be viewed cautiously, however, because only qualitative estimates of hybrid fitness are available for many animal species.

Overall, our data indicate that many plant species do reflect reproductively independent lineages ($\sim 70\%$) and thus appear to represent biologically real entities. However, we recognize that the CI is a fairly crude measure of reproductive independence because only postmating reproductive barriers were assayed. Species in some groups may be isolated entirely by premating barriers¹⁹, which would lead to an underestimate of congruence between species taxa and reproductive communities. This might explain the poor correspondence between species taxa and CI found in many animal groups (Fig. 2), where behaviour has an important role in reproductive isolation. More generally, our focus on postmating barriers implies that our estimates of correspondence between species taxa and reproductively independent lineages are conservative.

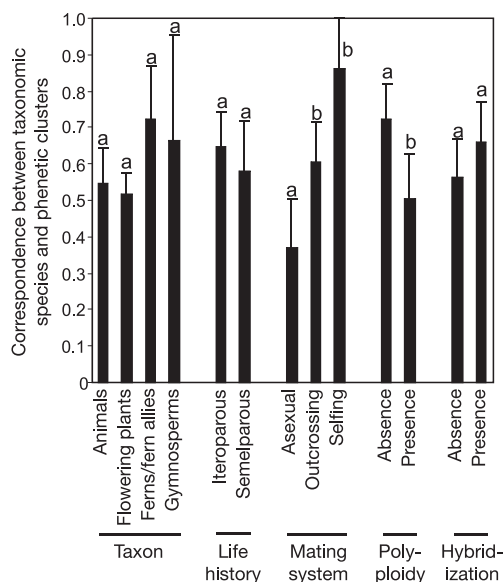


Figure 1 | Proportion of species taxa that correspond directly to phenotypic clusters compared on the basis of taxon, life history, mating system, polyploidy and contemporary hybridization. Refer also to Table 1. For each comparison, means with different letters are significantly different at $P < 0.05$ (Tukey's 'honestly significant difference' (HSD) test). Error bars indicate s.e.m.

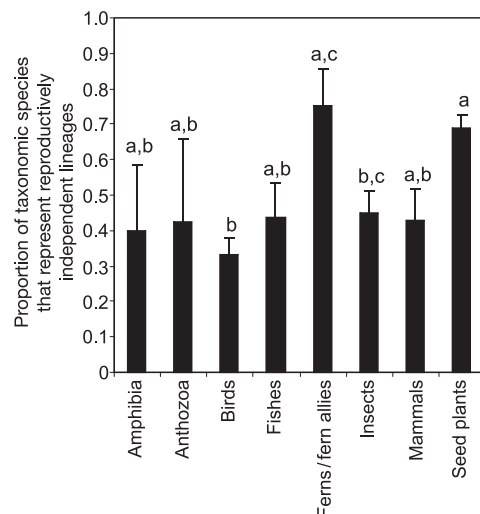


Figure 2 | Fraction of species taxa that represent reproductively independent lineages in major taxonomic groups of plants and animals. Means that do not share any letters are significantly different at $P < 0.05$ (Tukey's HSD test). Error bars indicate s.e.m.

Also, full fertility of intraspecific hybrids is not necessarily equivalent to the reproductive cohesion of conspecific populations. Indeed, botanists have frequently questioned whether there is sufficient gene flow among conspecific plant populations to allow them to evolve as an evolutionary unit^{1–3}. However, historical estimates of gene flow from molecular marker studies²⁰ imply that gene flow levels are much higher than suggested by earlier direct estimates²¹, and are roughly equivalent across plants and animals. Theoretical advances in population genetics further indicate that even low-migration species may evolve in concert through the spread of advantageous alleles²⁰ or through local extinction/colonization and source–sink dynamics in subdivided populations²². The latter two processes, which are common in natural populations of plants²³, lead to a reduction in effective population size and shortened species-wide fixation times for both beneficial and deleterious alleles²².

Botanists have been accused of poisoning Darwin's mind about the nature of species²⁴ and our results at least partly validate this accusation. Although the taxonomic problems caused by agamospermy in plants are real, apomixis has been reported for only 126 (<1%) of 13,000 genera of seed plants¹⁵. Thus, the large impact of agamic complexes on the psyche of botanists^{1–4,8} may have more to do with the abundance of certain agamic groups (for example, dandelions and blackberries), and fascination with them by taxonomists, than to their contribution to overall plant diversity. In the majority of sexual plant taxa, discrete entities that correspond to reproductively independent lineages do exist at the species level and a useful classification would reflect this.

METHODS

Phenetic analyses. To identify appropriate numerical taxonomic studies of plants and animals, we surveyed journals with an experimental taxonomic or evolutionary focus (*Plant Systematics and Evolution*, *Systematic Botany*, *Watsonia*, *Botanical Journal of the Linnean Society*, *Systematic Zoology/Biology and Journal of Zoology*), as well as *The Bryologist* and the *American Fern Journal* to increase representation from basal vascular plants. We also searched ISI's *Web of Science* using the search terms, "plant and phenetic" and "principal component and phenetic" and "morphometric and taxonomy".

For each taxon, we recorded the number of discrete (that is, non-overlapping) clusters identified by either ordination or clustering methods (Supplementary Tables 1 and 2). Taxa differentiated by discriminant analysis, but not by other phenetic approaches, were not considered to be discrete because discriminant analysis weights characters that are perceived to be of importance in discriminating previously recognized groups.

Multifactor ANOVA was used to test for significant effects of taxon, life history (iteroparous versus semelparous), mating system (asexual, outcrossing, selfing), polyploidy (absence versus presence) and contemporary hybridization (absence versus presence) on correspondence between species taxa and phenetic clusters. When feasible, taxonomic groups polymorphic for life history were subdivided into groups that were monomorphic for life history; otherwise, life history was treated as 'missing'. Likewise, taxa containing both selfing and outcrossing species were either subdivided into groups that were monomorphic for mating system, or mating system was treated as 'missing'. However, groups containing both sexual and asexual taxa were treated as 'asexual' in the analyses because asexual taxa often overlapped phenotypically with multiple sexual taxa. Multiple studies within a genus were pooled to avoid phylogenetic bias. Contrasts compared asexual versus sexual taxa and selfing versus outcrossing sexual taxa. All analyses were performed using SAS PROC GLM (SAS Institute, 2001). Because studies in some genera included few taxa, leading to a discontinuous distribution of correspondence values, we tested significance using 10,000 randomization tests as well as *F*-tests based on the untransformed percentages (see the Supplementary Methods). Similar results were obtained by both methods.

Crossability analyses. For plants, we screened *Plant Systematics and Evolution*, *Systematic Botany* and the *American Journal of Botany* (from 1960 on) for studies that reported on crossability from both intra- and inter-specific crosses (Supplementary Table 5). We also used ISI's *Web of Science* to search for crossability studies of the same genera used in the phenetic analyses (Supplementary Table 3). As before, genera that were polymorphic for important biological factors were subdivided into groups that were monomorphic for these factors for statistical analyses (see below). If intraspecific fertility data were unavailable for a group, then crossability indices (CIs) had to fall below 0.8 (the average significance

threshold) before significance was declared. Also, if multiple measures of reproductive isolation were provided, a multiplicative fitness function was used to calculate CI (calculation of CI based on the single-lowest factor yielded the same results; see the Supplementary Methods and Results).

Quantitative estimates of both intra- and inter-specific crossability are rarer in animals, so we relied mainly on several large compilations that provide an 'isolation index' between species of birds²⁵, flies²⁶, fishes²⁷, frogs²⁸ and lepidopterans²⁹ (Supplementary Table 6). For mammals, we extracted hybrid fitness data from Gray³⁰, whereas the remaining animal crossing studies were obtained from the primary literature following searches on ISI's *Web of Science*. To enable comparisons between the plant and animal data sets, the various isolation indices were converted to a crossability index. For flies²⁶, frogs²⁸ and lepidopterans²⁹, isolation indices that ranged from 0 (no isolation) to 1 (complete isolation) were inverted. For birds²⁵, fishes²⁷ and mammals³⁰, isolation indices were converted to a crossability index as follows: 1, both sexes fully fertile and viable; 0.75, one sex fertile, the other sex some individuals recorded as fertile; 0.5, one sex fertile, one sex viable but infertile; 0.25, one sex sometimes fertile, one sex viable but infertile; 0, both sexes either inviable or infertile or both²⁵. As before, if no information was provided on intraspecific compatibility, then the mean CI threshold of 0.8 was employed to assess correspondence. This classification maximizes estimates of the proportion of 'good' animal species (as compared to plant species); animal species are scored as reproductively independent lineages if their hybrids show any loss of fertility or viability.

One-way ANOVA was conducted to determine whether kingdom or class/division significantly affected correspondence between species taxa and CI. The analysis of the effects of kingdom was repeated with mean CI as the dependent variable. Although no significant departures from normality and equality of variances were observed, all analyses were performed with and without arcsine-square root transformations, with only the latter reported here.

Analysis of biological factors that might affect correspondence between species taxa and CI in plants was also conducted by ANOVA, with taxonomic division, life history, mating system and polyploidy included as main effects (see Supplementary Table 4). Mean correspondence was analysed with and without arcsine-square root transformations.

Received 26 July; accepted 3 November 2005.

- Levin, D. A. The nature of plant species. *Science* **204**, 381–384 (1979).
- Raven, P. H. in *Modern Aspects of Species* (eds Iwatsuki, K., Raven, P. H. & Bock, W. J.) 11–29 (Univ. of Tokyo Press, Tokyo, 1986).
- Hickman, J. C. *The Jepson Manual of Higher Plants in California* (Univ. of California Press, Berkeley, 1993).
- Bachmann, K. Species as units of diversity: an outdated concept. *Theory Biosci.* **117**, 213–230 (1998).
- Mishler, B. D. in *Species: New Interdisciplinary Essays* (ed. Wilson, R. A.) 307–316 (MIT Press, Cambridge, Massachusetts, 1999).
- Fisher, R. A. in *Evolution as a Process* (eds Huxley, J. S., Hardy, A. C. & Ford, E. B.) 84–98 (Allen and Unwin, London, 1954).
- Diamond, J. M. Horrible plant species. *Nature* **360**, 627–628 (1992).
- Grant, V. in *The Species Problem* (ed. Mayr, E.) 38–90 (American Association for the Advancement of Science, Washington DC, 1957).
- Mayr, E. A local flora and the biological species concept. *Am. J. Bot.* **79**, 222–238 (1992).
- McDade, L. A. Species concepts and problems in practice: Insight from botanical monographs. *Syst. Bot.* **20**, 606–622 (1995).
- Berlin, B., Breedlove, D. E. & Raven, P. H. *Principles of Tzeltal Plant Classification* (Academic, New York/London, 1974).
- Wang, J. X., Liu, H. M., Hu, H. B. & Gao, L. Participatory approach for rapid assessment of plant diversity through a folk classification system in a tropical rainforest: Case study in Xishuangbanna, China. *Conserv. Biol.* **18**, 1139–1142 (2004).
- Ridley, M. *Evolution* (Blackwell Scientific Publications, Oxford, 1996).
- Mishler, B. D. & Donoghue, M. J. Species concepts: A case for pluralism. *Syst. Zool.* **31**, 491–503 (1982).
- Sokal, R. R. & Sneath, P. H. A. *Principles of Numerical Taxonomy* (W.H. Freeman and Company, San Francisco, 1963).
- Baker, H. G. Race formation and reproductive method in flowering plants. *Symp. Soc. Exp. Biol.* **7**, 114–145 (1953).
- Carman, J. G. Asynchronous expression of duplicate genes in angiosperms may cause apomixis, bispory, tetraspory, and polyembryony. *Biol. J. Linn. Soc.* **61**, 51–94 (1997).
- McDade, L. A. & Lundberg, J. G. A new tabular and diagrammatic method for displaying artificial hybridization data, with an example from *Aphelandra* (Acanthaceae). *Syst. Bot.* **7**, 13–25 (1982).
- Arnold, M. L. *Natural Hybridization and Evolution* (Oxford Univ. Press, New York, 1997).
- Morjan, C. L. & Rieseberg, L. H. How species evolve collectively: implications of

- gene flow and selection for the spread of advantageous alleles. *Mol. Ecol.* **13**, 1341–1356 (2004).
21. Ehrlich, P. R. & Raven, P. H. Differentiation of populations. *Science* **165**, 1228–1232 (1969).
 22. Whitlock, M. C. Fixation probability and time in subdivided populations. *Genetics* **164**, 767–779 (2003).
 23. Levin, D. A. *The Origin, Expansion, and Demise of Plant Species* (Oxford Univ. Press, New York, 2000).
 24. Mayr, E. *The Growth of Biological Thought* (Harvard Univ. Press, Cambridge, Massachusetts, 1982).
 25. Price, T. D. & Bouvier, M. M. The evolution of F1 postzygotic incompatibilities in birds. *Evolution* **56**, 2083–2089 (2002).
 26. Coyne, J. A. & Orr, H. A. "Patterns of speciation in *Drosophila*" revisited. *Evolution* **51**, 295–303 (1997).
 27. Russell, S. T. Evolution of intrinsic post-zygotic reproductive isolation in fish. *Ann. Zool. Fenn.* **40**, 321–329 (2003).
 28. Sasa, M. M., Chippindale, P. T. & Johnson, N. A. Patterns of postzygotic isolation in frogs. *Evolution* **52**, 1811–1820 (1998).
 29. Presgraves, D. C. Patterns of postzygotic isolation in Lepidoptera. *Evolution* **56**, 1168–1183 (2002).
 30. Gray, A. P. *Mammalian Hybrids: A Check-list with Bibliography* (Commonwealth Agricultural Bureaux Farnham Royal, Slough, England, 1972).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Funding for this work was provided by the US National Institutes of Health, the US National Science Foundation, and the Guggenheim Foundation (to L.H.R.). We thank M. Barker, S. Bickford, A. Buerkle, J. Burke, B. Gross, M. Moody, J. Strasburg, Y. Supir, M. Tseng and K. Whitney for helpful comments on an earlier version of the manuscript, and the Centre for Plant Biodiversity Research at CSIRO Plant Industry, Australia for providing office space and library resources for this project. W. Hastie, M. Hearn and M. Thornton, from the CSIRO Black Mountain Library, provided extensive bibliographic assistance.

Author Contributions L.H.R. conceived of the literature surveys, T.E.W. and L.H.R. collected the data, and E.J.B. and L.H.R. conducted the data analyses. The manuscript was written by L.H.R. with comments and assistance from T.E.W. and E.J.B.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.H.R. (Iriesebe@indiana.edu).

LETTERS

The prolyl isomerase Pin1 regulates amyloid precursor protein processing and amyloid- β production

Lucia Pastorino^{1*}, Anyang Sun^{1*†}, Pei-Jung Lu^{2*}, Xiao Zhen Zhou¹, Martin Balastik¹, Greg Finn¹, Gerburg Wulf¹, Jormay Lim¹, Shi-Hua Li³, Xiaojiang Li³, Weiming Xia⁴, Linda K. Nicholson⁵ & Kun Ping Lu¹

Neuropathological hallmarks of Alzheimer's disease are neurofibrillary tangles composed of tau and neuritic plaques comprising amyloid- β peptides (A β) derived from amyloid precursor protein (APP), but their exact relationship remains elusive^{1–3}. Phosphorylation of tau and APP on certain serine or threonine residues preceding proline affects tangle formation and A β production *in vitro*^{3–5}. Phosphorylated Ser/Thr-Pro motifs in peptides can exist in *cis* or *trans* conformations, the conversion of which is catalysed by the Pin1 prolyl isomerase^{6,7}. Pin1 has been proposed to regulate protein function by accelerating conformational changes^{7–10}, but such activity has never been visualized and the biological and pathological significance of Pin1 substrate conformations is unknown⁷. Notably, Pin1 is downregulated and/or inhibited by oxidation in Alzheimer's disease neurons, *Pin1* knockout causes tauopathy and neurodegeneration^{8,9,11,12}, and *Pin1* promoter polymorphisms appear to associate with reduced Pin1 levels and increased risk for late-onset Alzheimer's disease^{13,14}. However, the role of Pin1 in APP processing and A β production is unknown. Here we show that Pin1 has profound effects on APP processing and A β production. We find that Pin1 binds to the phosphorylated Thr668-Pro motif in APP and accelerates its isomerization by over 1,000-fold, regulating the APP intracellular domain between two conformations, as visualized by NMR. Whereas Pin1 overexpression reduces A β secretion from cell cultures, knockout of *Pin1* increases its secretion. *Pin1* knockout alone or in combination with overexpression of mutant APP in mice increases amyloidogenic APP processing and selectively elevates insoluble A β 42 (a major toxic species) in brains in an age-dependent manner, with A β 42 being prominently localized to multivesicular bodies of neurons, as shown in Alzheimer's disease before plaque pathology¹⁵. Thus, Pin1-catalysed prolyl isomerization is a novel mechanism to regulate APP processing and A β production, and its deregulation may link both tangle and plaque pathologies. These findings provide new insight into the pathogenesis and treatment of Alzheimer's disease.

A primary theory for the cause of Alzheimer's disease is overproduction and/or lack of clearance of A β peptides derived from APP, especially the more toxic A β 42 (refs 1, 2). Phosphorylation of APP on the Thr 668-Pro motif is elevated in the brains of Alzheimer's disease sufferers, and increases A β secretion *in vitro*⁵, although its *in vivo* significance in regulating APP processing and A β production is unknown. Notably, after Thr 668 phosphorylation, a new

population of the pThr 668-Pro motif appears in the *cis* conformation and exchanges very slowly with the *trans* form^{16,17}. Because Thr 668 phosphorylation is known to be increased during mitosis¹⁸, like many other Pin1 substrates^{6–9}, and Pin1 is important for protecting against tauopathy and neurodegeneration^{7,8,11}, we hypothesize that Pin1 might act on the pThr668-Pro motif to regulate APP processing and A β production.

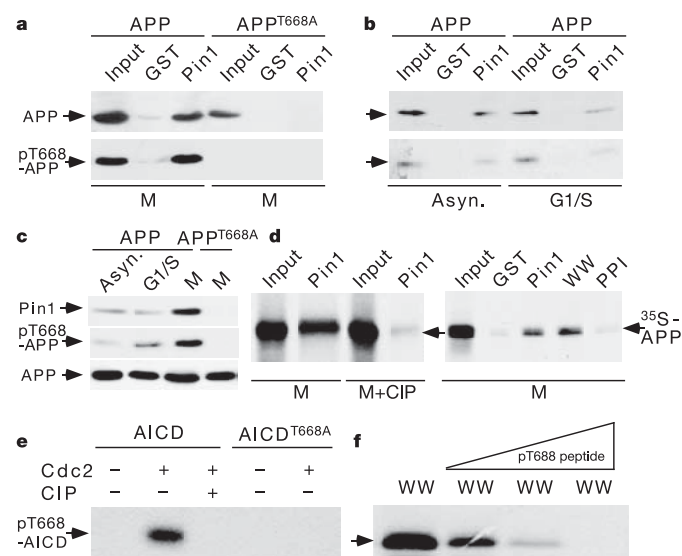


Figure 1 | Pin1 binds to the phosphorylated Thr 668-Pro motif in APP *in vitro* and *in vivo*. **a–c**, N18 neuroblastoma cells transfected with the HA-APP or HA-APP^{T668A} construct were arrested at M or G1/S boundary or left asynchronous (Asyn.), followed by GST pull down and immunoblot for APP and pT668-APP (**a**, **b**), or by anti-HA monoclonal antibody immunoprecipitation and then immunoblot for Pin1, APP and pT668-APP (**c**). **d**, ³⁵S-APP was phosphorylated by *Xenopus* mitotic extracts (M) or first phosphorylated by mitotic extracts and then dephosphorylated with CIP (M+CIP), followed by GST pull down. **e**, AICD and AICD^{T668A} phosphorylated by cyclin B/Cdc2 were subjected to GST pull down directly or after dephosphorylation with CIP, followed by immunoblot with anti-pT668-APP antibodies. **f**, ³²P-AICD was subjected to GST-WW pull down in the presence of increasing amounts of a pThr 668-containing APP phosphopeptide.

¹Cancer Biology Program, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215, USA. ²Department of Medical Education and Research, Kaohsiung Veterans General Hospital, Kaohsiung 804, Taiwan, China. ³Department of Human Genetics, Emory University, Atlanta, Georgia 30322, USA. ⁴Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁵Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. [†]Present address: Institute of Neurobiology, Fudan University, Shanghai 200433, China.

*These authors contributed equally to this work.

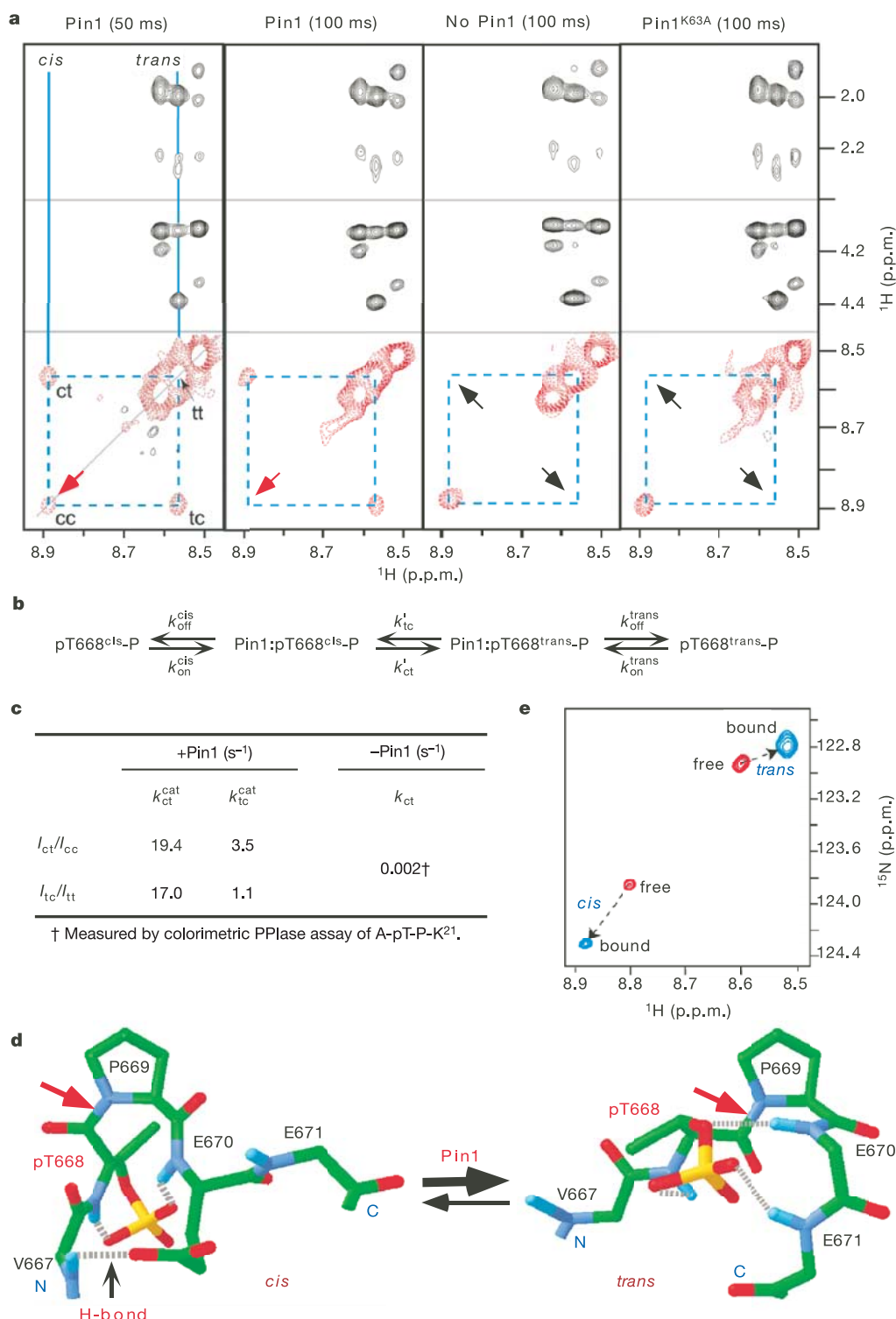


Figure 2 | Pin1 catalyses isomerization of the pThr 668-Pro motif in APP as visualized by NMR spectroscopy. a, Selected region of E670-¹H^N two-dimensional ROESY spectra of a pThr 668-Pro-containing APP peptide at 3 mM in the presence or absence of 0.05 mM GST-Pin1 or its K63A mutant at $t_m = 50$ or 100 ms. Negative (positive) intensity is denoted in red (black). E670-¹H^N diagonal peaks for *cis* (cc) and *trans* (tt) display exchange peaks (ct and tc) only in the presence of active Pin1 (left two panels). **b**, Enzyme kinetics scheme for Pin1-catalysed isomerization of the pThr 668-Pro bond. **c**, Pin1-catalysed *cis* to *trans* ($k_{\text{ct}}^{\text{cat}}$) and *trans* to *cis* ($k_{\text{tc}}^{\text{cat}}$) isomerization rates from independent fitting of E670-¹H^N *cis* and *trans* ROESY peak intensity

ratios ($I_{\text{ct}}/I_{\text{cc}}$ and $I_{\text{tc}}/I_{\text{tt}}$ respectively, where I_{jk} denotes intensity of peak $jk = \text{ct, cc, tc, tt}$), with the uncatalysed *cis* to *trans* isomerization rate (k_{ct}) of a pThr-Pro peptide²¹ for comparison. **d**, Pin1-catalysed isomerization between *cis* and *trans* conformations of the pThr 668-Pro peptide bond (red arrows). Structural models display local backbone conformations with associated hydrogen bonds. The size and direction of black arrows represent the catalysis reaction accelerated by Pin1. **e**, Regions extracted from the two-dimensional ¹⁵N-¹H HSQC spectrum of 0.2 mM ¹⁵N-E670-pT668 peptide in the absence (red) and presence of equimolar amounts of WW domain (blue).

To test this hypothesis, we first examined whether Pin1 interacts with APP by transfecting N18 cells with a HA-APP construct, followed by arresting them at mitosis or G1/S to manipulate Thr 668 phosphorylation of APP before glutathione S-transferase (GST) pull down and co-immunoprecipitation^{6,8,19}. Pin1 bound to expressed and endogenous APP from mitotic cells (Fig. 1a; see also Supplementary Fig. S1a) and to a lesser extent from asynchronous or G1/S cells (Fig. 1b). Similarly, anti-HA monoclonal antibody co-immunoprecipitated endogenous Pin1 from mitotic cells and to a lesser degree from asynchronous or G1/S cells (Fig. 1c). These differences in Pin1 binding to APP correlated with cell-cycle-regulated Thr 668 phosphorylation (Fig. 1a–c)¹⁸. To examine the importance of APP phosphorylation for Pin1 binding, ³⁵S-APP was synthesized and phosphorylated by mitotic extracts, followed by dephosphorylation before GST pull down^{6,8}. Pin1 bound to phosphorylated APP; the binding was abolished by dephosphorylation and mediated by the WW domain of Pin1 (Fig. 1d), a known pSer/Thr-Pro-binding module¹⁹.

To identify the Pin1-binding site in APP, we substituted Ala for Thr 668, the only Thr/Ser-Pro motif in the APP intracellular domain (AICD). Pin1 failed to bind to the APP^{T668A} mutant when expressed in mitotic cells (Fig. 1a, c), indicating that the Thr 668-Pro motif is important for binding. To confirm these results, we incubated recombinant AICD and AICD^{T668A} with cyclin B/Cdc2 kinase, followed by dephosphorylation with CIP before GST pull down. As shown¹⁸, AICD, but not AICD^{T668A}, was phosphorylated by Cdc2 (Supplementary Fig. S1b). Notably, Pin1 bound to Cdc2-phosphorylated AICD, but not AICD^{T668A}, and the binding was abolished by dephosphorylation (Fig. 1e). This binding was mediated by the WW domain of Pin1 (Supplementary Fig. S1c) and effectively competed by a pThr 668-containing APP peptide (Fig. 1f). These results indicate that Pin1 binds to the pThr 668-Pro motif in APP *in vitro* and *in vivo* and that the binding is mediated by the WW domain, as is the case for other Pin1 substrates^{4,7,19}.

Pin1 is shown to catalyse *cis/trans* isomerization of pSer/Thr-Pro motifs using biochemical assays^{6,9,10}, but such activity has never been visualized with atomic resolution. Here we used NMR spectroscopy to examine Pin1-catalysed isomerization of the pThr 668-Pro bond in a 21-residue phosphopeptide (G659–Q679) of APP, because this peptide accurately reflects the structural and dynamic features of the corresponding residues in the native AICD¹⁶. Slow exchange between pThr 668-Pro isomers yields two distinct sets of ¹H peaks for many residues in the peptide, with the E670 amide proton (E670-H^N) particularly well resolved (Fig. 2a). In the ROESY (rotating frame Overhauser effect spectroscopy) spectrum, the intensities of exchange and diagonal peaks for a two-state isomerization reaction explicitly depend on the forward and backward rate constants (Fig. 2b, k_{ct}^{cat} and k_{tc}^{cat}), and on the mixing time (t_m). With increasing t_m , diagonal peaks diminished and exchange peaks grew (Fig. 2a, red arrows). The same exchange cross peaks were observed upon the addition of a catalytic amount of GST–Pin1 or Pin1; however, exchange cross peaks were absent when either no Pin1 or a catalytically inactive Pin1^{K63A} mutant was added (Fig. 2a, black arrows). Ratios of cross/diagonal peak intensities for *cis* and *trans* E670-H^N were calculated (Supplementary Fig. S2), as described²⁰, and curve fits of these ratios against t_m provided independent and consistent measures of k_{ct}^{cat} and k_{tc}^{cat} (Fig. 2c; see also Supplementary Fig. S2). Pin1 accelerated the pThr 668-Pro isomerization rate by several orders of magnitude over the typical uncatalysed isomerization rates for pThr-Pro peptides²¹ and markedly reduced the average lifetime of the *cis* (~0.05 s) and *trans* (~0.5 s) isomeric states (Fig. 2c, d). The k_{ct}^{cat} and k_{tc}^{cat} rates differ by tenfold, as expected based on the equilibrium populations of free *cis* and *trans*. These results provide a direct atomic-level demonstration of Pin1-catalysed conformational regulation of its substrates.

Given that the Pin1 WW domain mediates substrate recognition (Fig. 1), the question arises whether differences in WW domain

affinity for the *cis* and *trans* isomers might alter their relative populations when Pin1 and pThr 668-APP concentrations are comparable. To address this question, we compared ¹⁵N-¹H correlation peak volumes corresponding to free *cis* and *trans* and *cis* and *trans* bound to the isolated Pin1 WW domain using the 21-residue pThr 668 peptide with ¹⁵N incorporated at E670. Addition of the WW domain caused significant peak shifts for both *cis* and *trans* isomers (Fig. 2e), demonstrating direct binding between the WW domain and both pThr 668-Pro isomers. However, the population distribution remained similar—13% *cis* and 87% *trans*—after binding (Fig. 2e), indicating that the WW domain exhibits comparable affinity for *cis* and *trans* pThr 668-Pro motifs in APP. Because Pin1 WW domain is shown to bind to other phosphopeptides in *trans*²², these results further suggest that Pin1 WW binding to the *cis* conformation is highly sequence-specific.

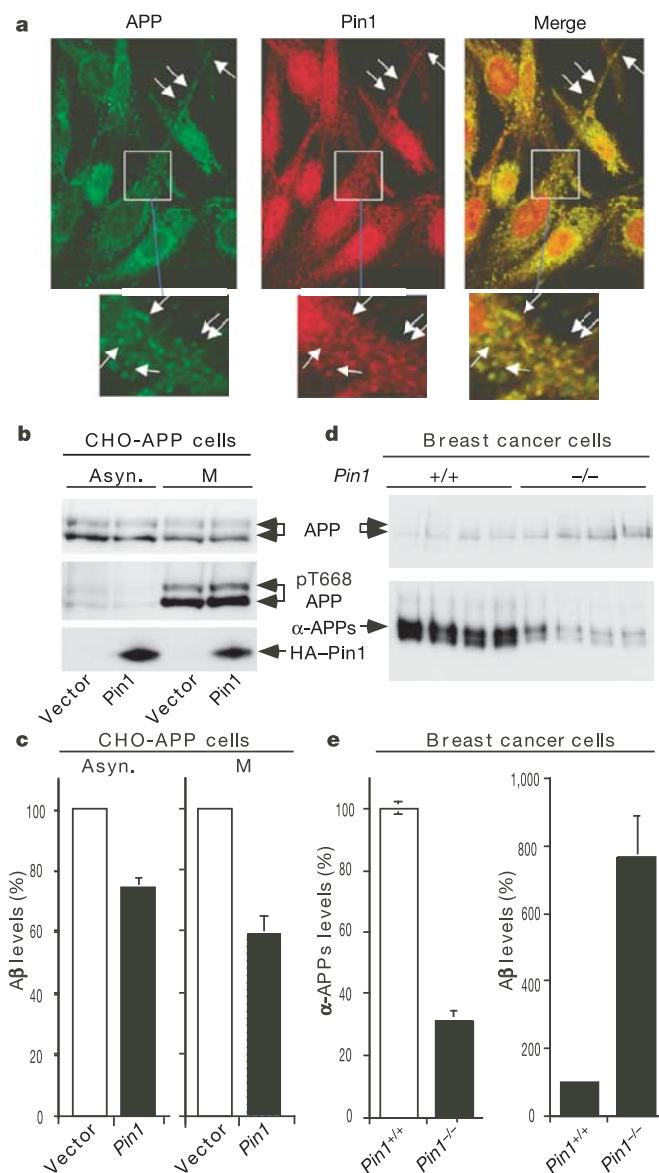


Figure 3 | Pin1 co-localizes with APP and regulates APP processing and Aβ secretion in cells. **a**, CHO-APP cells were doubly immunostained for Pin1 and APP, followed by confocal microscopy. **b**, **c**, CHO-APP cells transfected with Pin1 construct or vector were left asynchronous or arrested at M, followed by immunoblot for Pin1, APP and pT668-APP (**b**) or ELISA measuring total Aβ secretion (**c**). **d**, **e**, After Pin1^{+/+} and Pin1^{-/-} breast cancer cells were transfected with an APP construct, secreted α-APPs were assayed by immunoprecipitation and then immunoblot (**d**), followed by semi-quantification and total Aβ by ELISA (**e**). Error bars, s.d.

Given that Pin1 binds to the pThr668-Pro motif in APP and greatly accelerates its isomerization, a critical question is whether Pin1 affects APP processing and A β production. APP is processed by non-amyloidogenic α -secretases mainly at the plasma membrane, and by amyloidogenic β - and γ -secretases at endosomes and other structures^{1,2}. Therefore, it is of great importance to examine whether and where Pin1 and APP co-localize in the cell. We found that Pin1 and APP co-localized prominently at the plasma membrane and in intracellular vesicles close to the plasma membrane both in CHO-APP (Chinese hamster ovary cell line stably expressing the wild-type human APP751 isoform) and H4 cells (Fig. 3a; see also Supplementary Fig. S3). These vesicles were further identified to be clathrin-coated vesicles but not endosomes or subsequent structures (Supplementary Figs S4–S6). Therefore, Pin1 co-localizes with APP primarily at the plasma membrane and clathrin-coated vesicles, suggesting that Pin1 might promote non-amyloidogenic APP processing and reduce A β production.

To examine this possibility, we first determined the effects of Pin1 overexpression on A β secretion in cultured cells by transfecting CHO-APP cells with a *Pin1* construct or CHO cells with *Pin1* and *APP751* constructs, followed by measuring total A β secretion from

asynchronous or mitotic cells. Pin1 overexpression in CHO-APP and CHO cells significantly reduced A β secretion, especially from mitotic cells where Thr 668 phosphorylation was elevated (Fig. 3b, c; see also Supplementary Fig. S7). These results show that overexpression of Pin1 reduces A β secretion and the effects might depend on Thr 668 phosphorylation.

We next examined the effects of *Pin1* knockout on APP processing and A β secretion using breast cancer cell lines derived from mouse mammary tumour virus (MMTV)-Neu transgenic mice in a *Pin1*^{+/+} or *Pin1*^{-/-} background, as described²³. APP is cleaved by α - or β -secretases to generate soluble amino-terminal fragments (α -APPs or β -APPs) and membrane-anchored carboxy-terminal fragments (α -CTFs or β -CTFs), respectively^{1,2,24}. *Pin1*^{+/+} and *Pin1*^{-/-} cancer cell lines expressed comparable levels of endogenous APP, α -CTFs and β -CTFs (Supplementary Fig. S8a) or human APP after transfection (Fig. 3d). However, as compared with *Pin1*^{+/+} cells, *Pin1*^{-/-} cells secreted ~3-fold less α -APPs (Fig. 3d, e), as detected by 6E10 specific for human α -APPs (Supplementary Fig. S8b), but ~7-fold more A β (Fig. 3e). These results indicate that *Pin1* knockout in cells decreases α -APPs secretion, but increases A β secretion.

To examine whether *Pin1* knockout affects APP processing and A β

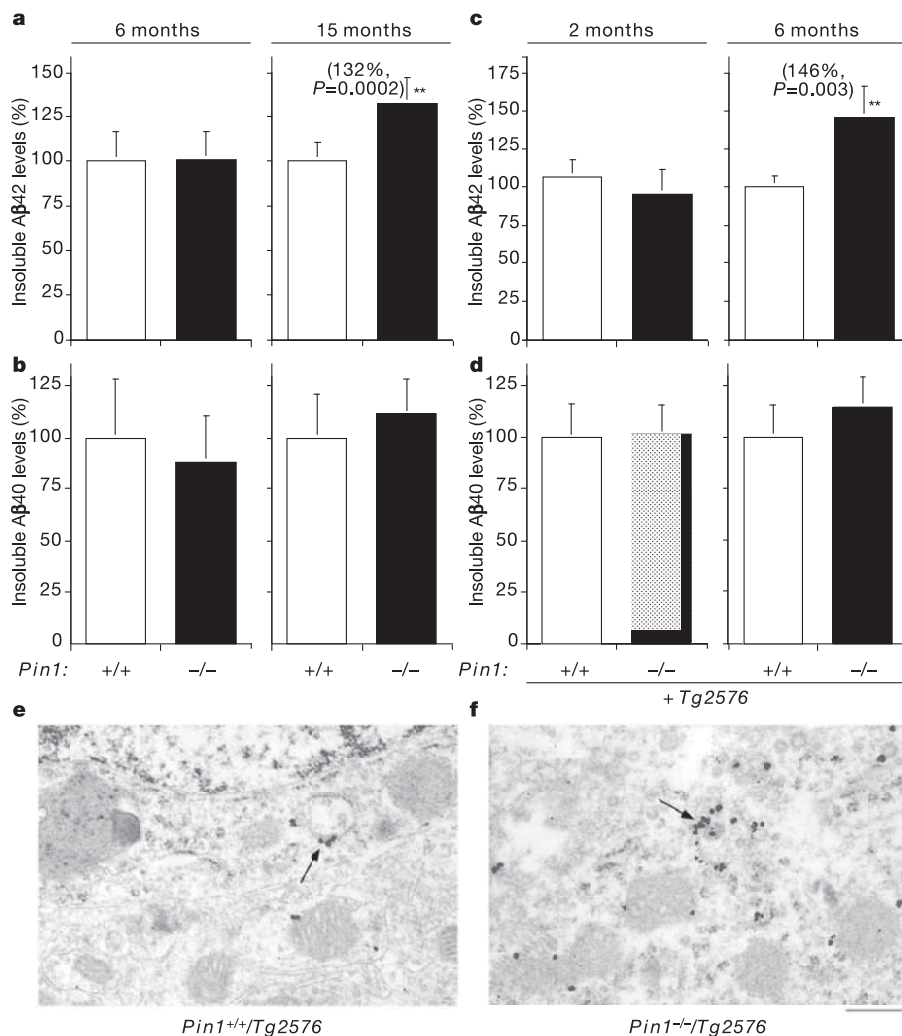


Figure 4 | *Pin1* knockout causes age-dependent and selective accumulation of insoluble A β 42 at multivesicular bodies of neurons, which is accelerated by APP mutant overexpression. **a–d**, Brain tissues were collected from different ages of *Pin1*^{-/-} and *Pin1*^{+/+} littermates (**a**, **b**) or *Pin1*^{-/-} and *Pin1*^{+/+} littermates in the presence of transgenic overexpression of a single copy of *APP*^{K670N/M671L} (Tg2576) (**c**, **d**). Soluble and insoluble A β peptides were extracted from brain tissues and A β 40 and A β 42 levels were measured

by ELISA and are represented as means $\pm$ s.d. **e**, **f**, Tg2576 littermates in a *Pin1*^{+/+} (**e**) or *Pin1*^{-/-} (**f**) background were perfused at 7 months old and processed for immunogold electron microscopy using anti-human A β 42 antibodies. Immunogold particles of A β 42 are primarily localized to multivesicular bodies (arrows) in dorsal medial cortical neurons. Scale bar, 500 nm.

production in mice, we first examined A β production in the brains of *Pin1*^{-/-} mice using an enzyme-linked immunosorbent assay (ELISA), as described²⁵. Sequential proteolysis of APP by β - and γ -secretases generates mainly 40- and 42-residue A β peptides (A β 40 and A β 42), with A β 42 being the major toxic species and key contributor to plaque formation in Alzheimer's disease^{1,2,25-29}. Because the effects of *Pin1* knockout on neurons in mice are age dependent¹¹, we compared A β levels at 2–6 months of age, when there is no detectable neuronal phenotypes, and at 15 months, when tau hyperphosphorylation and early neurodegeneration are observed¹¹. *Pin1* knockout did not significantly change the levels of A β 40 or A β 42 at 2–6 months of age (Fig. 4a, b; see also Supplementary Fig. S9a, b), or soluble A β 40, A β 42 and insoluble A β 40 at 15 months of age (Fig. 4b; see also Supplementary Fig. S9a, b). However, levels of insoluble A β 42 were increased by 32% in the brains of *Pin1*^{-/-} mice at 15 months compared to *Pin1*^{+/+} littermates (Fig. 4a).

To insure that the A β 42 increase is not due to early neurodegeneration in *Pin1*^{-/-} mice and to examine the effects of *Pin1* knockout on APP processing *in vivo*, we crossed *Pin1*^{-/-} mice with *Tg(APP^{SWE})2576Kha* mice (also called *Tg(APP^{SWE})2576Kha*s; hereafter called *Tg2576* mice) overexpressing the familial Alzheimer's disease (FAD) *APP*^{K670N/M671L} mutant²⁷. Knockout of *Pin1* did not obviously affect A β levels at 2 months of age, but significantly increased insoluble A β , especially A β 42, by 46% compared with *Pin1*^{+/+} littermates at 6 months without affecting soluble A β 40 and A β 42 or causing neurodegeneration (Fig. 4c, d; see also Supplementary Fig. S9c, d; data not shown). Immunogold electron microscopy studies showed that A β 42 primarily accumulated in multivesicular bodies of neurons from *Tg2576* littermates with or without *Pin1* at 7 months of age, but more gold particles were observed in the absence than the presence of *Pin1* (Fig. 4e, f: 1–5 versus >8 particles per multivesicular body). These results demonstrate that the increased A β 42 is prominently localized to multivesicular bodies, where A β 42

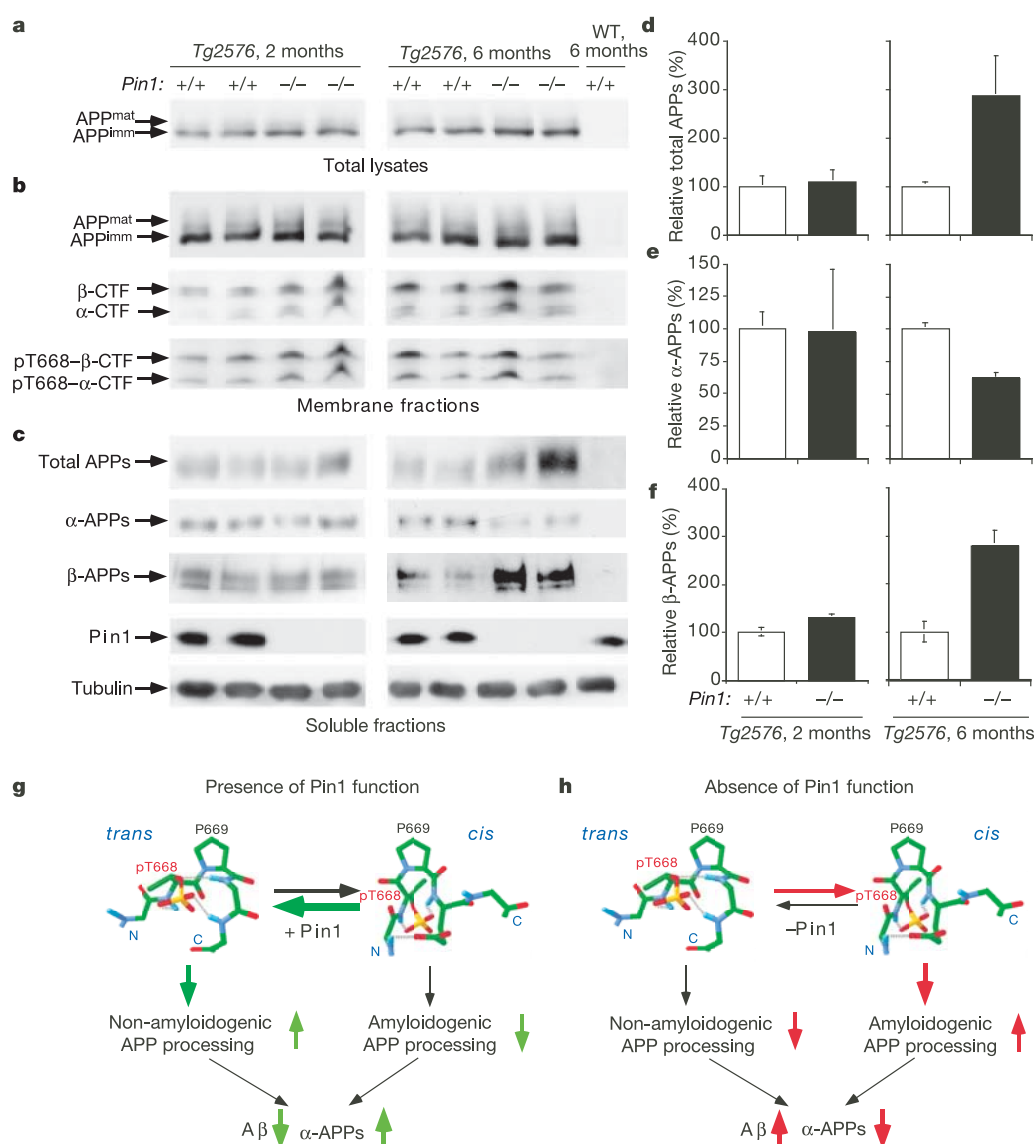


Figure 5 | *Pin1* knockout affects APP processing in mice in an age-dependent manner, and a model. **a–c**, Brain lysates (**a**) from 2- and 6-month-old *Tg2576* mice in the *Pin1*^{+/+} or *Pin1*^{-/-} background and non-transgenic controls were sub-fractionated into membrane fraction (**b**) and soluble fraction (**c**), followed by immunoblot with various specific antibodies. **d–f**, The changes of total APPs (**d**), α -APPs (**e**) and β -APPs (**f**) were semi-quantified with NIH Image and presented with *Pin1*^{+/+} controls

as 100% means = $\pm$ s.d. **g, h**, Although the pThr 668-Pro motif of APP tends to be in *cis* after phosphorylation, functional Pin1 would greatly accelerate *cis* to *trans* isomerization, which might favour non-amyloidogenic APP processing (**g**). Without proper Pin1 function, the *cis* pThr 668-Pro motif would not be isomerized to *trans* in a timely manner, which might favour amyloidogenic APP processing (**h**).

is known to be in the brains of human Alzheimer's disease sufferers and in *Tg2576* mouse brain before β -amyloid plaque pathology¹⁵. These results indicate that *Pin1* knockout in mice causes an age-dependent and selective increase in insoluble A β 42, which is accelerated by APP overexpression. Of note, similar increases in A β 42 levels have been documented in transgenic mice overexpressing FAD presenilin mutants^{25,28,29}, or in FAD brains²⁶, indicating that small shifts in A β 42 production may have an important impact on the development of Alzheimer's disease.

The above results indicate that *Pin1* knockout increases A β 42 production, suggesting that it might favour amyloidogenic APP processing. To investigate this possibility, we examined changes in APP processing in *Tg2576* mice in the presence or absence of Pin1 at 2 and 6 months, as described²⁴. For *Tg2576* mice, *Pin1* knockout had no obvious effects on total APP, CTFs or their Thr 668 phosphorylation, but significantly affected levels of soluble APPs at 6 months but not at 2 months old (Fig. 5a–f). Levels of total APPs and β -APPs were increased by ~3-fold, but levels of α -APPs were reduced by ~50% in brains of *Pin1*^{-/-} mice as compared with *Pin1*^{+/+} controls (Fig. 5a–f). These results indicate that *Pin1* knockout increases amyloidogenic versus non-amyloidogenic APP processing and elevates A β 42 in an age-dependent manner.

Our results identify Pin1-catalysed prolyl isomerization as a novel mechanism to regulate APP processing and A β production relevant to Alzheimer's disease. Thr 668-Pro phosphorylation may act as a conformational switch by increasing the *cis* content from ~0% to ~10%¹⁶. In the absence of a catalyst, *cis* to *trans* isomerization may present a major rate-limiting step for biological processes because the distinct isomer structures probably interact with different cellular proteins. Given the dramatic effects of Pin1 on APP conformational dynamics *in vitro* and on APP processing and A β production in cell cultures and animals, we hypothesize that the *cis* pThr 668-Pro conformation may favour amyloidogenic APP processing, whereas the *trans* conformation may favour non-amyloidogenic APP processing, and by catalysing their conversion, Pin1 promotes non-amyloidogenic APP processing and reduces A β production (Fig. 5g, h).

In this model, although APP is probably phosphorylated on the Thr 668-Pro motif in *trans*⁷, this motif partitions into both *trans* and *cis* isomers due to local structural constraints after phosphorylation^{16,17}. Pin1 would rapidly re-establish equilibrium if the *trans* (or *cis*) population were suddenly depleted. In the non-equilibrium cellular environment (Fig. 5g), Pin1-catalysed prolyl isomerization might prevent an increase in amyloidogenic APP processing and A β production by dynamically linking *cis* and *trans* populations and preventing the build up of *cis*. However, if Pin1 function is absent as in *Pin1*^{-/-} mice or cells, or downregulated/inhibited by oxidation or genetic alterations as in Alzheimer's disease (Fig. 5h)^{8,9,11–13}, a higher concentration of the *cis* pThr 668-Pro motif would be present for a longer time due to the extremely slow uncatalysed isomerization rate, which might promote amyloidogenic APP processing. Interestingly, Pin1 also binds to and isomerizes the pThr 231-Pro motif in tau^{8,9} and its knockout also causes an age-dependent accumulation of the pThr 231-Pro motif in the Alzheimer's disease tangle-specific conformation that is likely also in *cis*¹¹. Therefore, Pin1 deregulation might have similar conformational effects on specific pThr-Pro motifs in APP and tau, although they might lead to different pathological changes with a similar neurodegenerative outcome^{8,9,11}. In addition, Pin1 overexpression has an important role in cancer^{7,23,30}. Therefore, further studies on Pin1-catalysed conformational regulation and its biological and pathological significance should help elucidate the molecular events leading to Alzheimer's disease and cancer, and might also lead to the development of new therapies.

METHODS

The interaction and co-localization of Pin1 and APP were determined as

described^{16,8,11,19}. NMR analysis was performed as reported¹⁶. Briefly, for ROESY experiments, peptide was analysed in the absence or presence of Pin1, GST–Pin1 or its K63A mutant at 60:1 molar ratio (3 mM pThr 668-Pro and 0.05 mM Pin1 or its mutant). For ¹⁵N–¹H HSQC experiments, ¹⁵N-E670-pThr 668 peptide was dissolved in the same buffer adjusted to pH 6.7 and incubated with an equimolar amount of WW domain (0.2 mM). *Pin1*^{-/-} mice were inbred in mixed 129/Sv and C57L/B6 (ref. 11) mice, and crossed with *Tg2576* mice overexpressing the human APP^{K670N/M671L} (Swedish) mutant²⁷ (Taconic) to generate a single copy of the APP transgene in the *Pin1*^{+/+} and *Pin1*^{-/-} genetic background. APP processing was determined, as described²⁴, while A β 40 and A β 42 levels were measured by sandwich ELISA, as described²⁵. Subcellular localization of neuronal A β 42 was determined using immunogold electron microscopy with anti-human A β 42 polyclonal antibodies (Chemicon), as described^{11,15}. Full details of these methods are given in the Supplementary Information.

Received 21 October; accepted 12 December 2005.

- Mattson, M. P. Pathways towards and away from Alzheimer's disease. *Nature* **430**, 631–639 (2004).
- Hardy, J. & Selkoe, D. J. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* **297**, 353–356 (2002).
- Lee, V. M., Goedert, M. & Trojanowski, J. Q. Neurodegenerative tauopathies. *Annu. Rev. Neurosci.* **24**, 1121–1159 (2001).
- Lu, K. P., Liou, Y. C. & Vincent, I. Proline-directed phosphorylation and isomerization in mitotic regulation and in Alzheimer's disease. *Bioessays* **25**, 174–181 (2003).
- Lee, M. S. *et al.* APP processing is regulated by cytoplasmic phosphorylation. *J. Cell Biol.* **163**, 83–95 (2003).
- Yaffe, M. B. *et al.* Sequence-specific and phosphorylation-dependent proline isomerization: A potential mitotic regulatory mechanism. *Science* **278**, 1957–1960 (1997).
- Lu, K. P. Pinning down cell signaling, cancer and Alzheimer's disease. *Trends Biochem. Sci.* **29**, 200–209 (2004).
- Lu, P. J., Wulf, G., Zhou, X. Z., Davies, P. & Lu, K. P. The prolyl isomerase Pin1 restores the function of Alzheimer-associated phosphorylated tau protein. *Nature* **399**, 784–788 (1999).
- Zhou, X. Z. *et al.* Pin1-dependent prolyl isomerization regulates dephosphorylation of Cdc25C and tau proteins. *Mol. Cell* **6**, 873–883 (2000).
- Stukenberg, P. T. & Kirschner, M. W. Pin1 acts catalytically to promote a conformational change in Cdc25. *Mol. Cell* **7**, 1071–1083 (2001).
- Liou, Y.-C. *et al.* Role of the prolyl isomerase Pin1 in protecting against age-dependent neurodegeneration. *Nature* **424**, 556–561 (2003).
- Sultana, R. *et al.* Oxidative modification and down-regulation of Pin1 in Alzheimer's disease hippocampus: A redox proteomics analysis. *Neurobiol. Aging* ahead of print publication, 9 June 2005 (doi:10.1016/j.neurobiolaging.2005.05.005).
- Segat, L. *et al.* Pin1 promoter polymorphisms are associated with Alzheimer's disease. *Neurobiol. Aging* ahead of print publication, 26 December 2005 (doi:10.1016/j.neurobiolaging.2005.11.009).
- Wijsman, E. M. *et al.* Evidence for a novel late-onset Alzheimer disease locus on chromosome 19p13.2. *Am. J. Hum. Genet.* **75**, 398–409 (2004).
- Takahashi, R. H. *et al.* Intraneuronal Alzheimer A β 42 accumulates in multivesicular bodies and is associated with synaptic pathology. *Am. J. Pathol.* **161**, 1869–1879 (2002).
- Ramelot, T. A. & Nicholson, L. K. Phosphorylation-induced structural changes in the amyloid precursor protein cytoplasmic tail detected by NMR. *J. Mol. Biol.* **307**, 871–884 (2001).
- Ramelot, T. A., Gentile, L. N. & Nicholson, L. K. Transient structure of the amyloid precursor protein cytoplasmic tail indicates preordering of structure for binding to cytosolic factors. *Biochemistry* **39**, 2714–2725 (2000).
- Suzuki, T. *et al.* Cell cycle-dependent regulation of the phosphorylation and metabolism of the Alzheimer amyloid precursor protein. *EMBO J.* **13**, 1114–1122 (1994).
- Lu, P. J., Zhou, X. Z., Shen, M. & Lu, K. P. A function of WW domains as phosphoserine- or phosphothreonine-binding modules. *Science* **283**, 1325–1328 (1999).
- Ernst, R. R., Bodenhausen, G. & Wokaun, A. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Clarendon, Oxford, 1987).
- Schutkowski, M. *et al.* Role of phosphorylation in determining the backbone dynamics of the serine/threonine-proline motif and Pin1 substrate recognition. *Biochemistry* **37**, 5566–5575 (1998).
- Wintjens, R. *et al.* ¹H NMR study on the binding of Pin1 Trp-Trp domain with phosphothreonine peptides. *J. Biol. Chem.* **276**, 25150–25156 (2001).
- Wulf, G., Garg, P., Liou, Y. C., Iglehart, D. & Lu, K. P. Modeling breast cancer *in vivo* and *ex vivo* reveals an essential role of Pin1 in tumorigenesis. *EMBO J.* **23**, 3397–3407 (2004).
- Pastorino, L. *et al.* BACE (β -secretase) modulates the processing of APLP2 *in vivo*. *Mol. Cell. Neurosci.* **25**, 642–649 (2004).
- Citron, M. *et al.* Mutant presenilins of Alzheimer's disease increase production

- of 42-residue amyloid β -protein in both transfected cells and transgenic mice. *Nature Med.* **3**, 67–72 (1997).
26. Scheuner, D. *et al.* Secreted amyloid β -protein similar to that in the senile plaques of Alzheimer's disease is increased *in vivo* by the presenilin 1 and 2 and APP mutations linked to familial Alzheimer's disease. *Nature Med.* **2**, 864–870 (1996).
27. Hsiao, K. *et al.* Correlative memory deficits, A β elevation, and amyloid plaques in transgenic mice. *Science* **274**, 99–102 (1996).
28. Duff, K. *et al.* Increased amyloid- β 42(43) in brains of mice expressing mutant presenilin 1. *Nature* **383**, 710–713 (1996).
29. Borchelt, D. R. *et al.* Familial Alzheimer's disease-linked presenilin 1 variants elevate A β 1–42/1–40 ratio *in vitro* and *in vivo*. *Neuron* **17**, 1005–1013 (1996).
30. Suizu, F., Ryo, A., Wulf, G., Lim, J. & Lu, K. P. Pin1 regulates centrosome duplication and its overexpression induces centrosome amplification, chromosome instability and oncogenesis. *Mol. Cell. Biol.* **26**, 1463–1479 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Selkoe for advice, B. Arosio and G. Annoni for sharing results on Pin1 genetic analysis in Alzheimer's disease before publication, D. Selkoe, D. Schenk, P. Seubert, D. Goldgaber, E. Koo, J. Wang and H. Xu for reagents, and J. Sears, H. Yi and M. Wotkowicz for technical assistance. J.L. is a Fellow of the Human Frontier Research Program; G.W. is a NIH Mentored Clinician Scientist. This study was supported by a Taiwan NSC grant to P.J.L., an NSF grant to L.K.N. and NIH grants to X.L., W.X., X.Z.Z. and K.P.L.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.P.L. (klu@bidmc.harvard.edu).

A P-type ATPase required for rice blast disease and induction of host resistance

Martin J. Gilbert¹, Christopher R. Thornton¹, Gavin E. Wakley¹ & Nicholas J. Talbot¹

To cause diseases in plants, pathogenic microorganisms have evolved mechanisms to deliver proteins directly into plant cells, where they suppress plant defences and facilitate tissue invasion^{1–3}. How plant pathogenic fungi, which cause many of the world's most serious plant diseases, deliver proteins during plant infection is currently unknown. Here we report the characterization of a P-type ATPase-encoding gene, *MgAPT2*, in the economically important rice blast pathogen *Magnaporthe grisea*, which is required for exocytosis during plant infection. Targeted gene replacement showed that *MgAPT2* is required for both foliar and root infection by the fungus, and for the rapid induction of host defence responses in an incompatible reaction. Δ *Mgapt2* mutants are impaired in the secretion of a range of extracellular enzymes and accumulate abnormal Golgi-like cisternae. However, the loss of *MgAPT2* does not significantly affect hyphal growth or sporulation, indicating that the establishment of rice blast disease involves the use of *MgApt2*-dependent exocytotic processes that operate during plant infection.

Plant pathogenic fungi are able to breach the external surfaces of plants and proliferate rapidly within plant tissue⁴. The rice blast fungus *M. grisea* produces specialized infection structures, called appressoria, to gain entry to the rice leaf, and subsequently develops a network of invasive hyphae to colonize host tissue^{5,6}. The fungus grows extensively in host cells before causing disease symptoms within 3–4 days of infection⁵. *M. grisea* can also infect roots, and elaborates alternative infection structures that allow hyphae to invade root tissue and the vascular system to cause systemic plant infection⁷.

The rapid internalization and spread of *M. grisea* hyphae through plant tissue occurs before disease symptoms become apparent, indicating that the fungus might have a capacity for the evasion or suppression of plant defences⁵. Sequencing of the *M. grisea* genome also revealed that the fungus produces many secreted proteins, indicating that the deployment of a large armoury of secreted peptides might be an important aspect of its pathogenicity⁸. To investigate the biology of plant tissue colonization by *M. grisea*, we decided to focus attention on the operation of the Golgi apparatus and protein trafficking in the fungus. Here we describe the characterization of a gene encoding an integral membrane P-type ATPase from *M. grisea*, belonging to the type IV, Drs2 family of aminophospholipid translocases (APTs). These enzymes are required in *Saccharomyces cerevisiae* for efficient Golgi function and are involved in both endocytosis and exocytosis^{9–11}. APTs maintain the asymmetrical distribution of aminophospholipids in cellular membranes, in which the aminophospholipids phosphatidylserine and phosphatidylethanolamine are enriched within the inner leaflet of the membrane bilayer, and phospholipids such as sphingomyelin and glycosphingolipids are found predominantly in the outer leaflet^{10,11}. Phospholipid asymmetry is important in membrane fusion events, such as vesicle budding and docking, which take place at both the

plasma membrane and in the *trans*-Golgi network¹⁰. *M. grisea* possesses four putative APT-encoding genes, *MgPDE1*, *MgAPT2*, *MgAPT3* and *MgAPT4* (Fig. 1a). Phylogenetic analysis revealed the *MgAPT2* gene product (*M. grisea* genome database accession number MG02767, Broad Institute, USA) to be closely related to P-type ATPases of the Drs2 family of APTs (Fig. 1a and Supplementary Table 1). The *MgAPT2*-encoded protein consists of 1,524 amino acids with a predicted molecular mass of 171.3 kDa, has ten predicted membrane-spanning domains, and all conserved features expected of a P-type ATPase¹² (Fig. 1b). To determine whether *MgAPT2* is functionally related to the Drs2 family of APTs, complementation experiments were performed. In yeast, a Δ *dnf1* Δ *dnf2* Δ *dnf3* yeast mutant exhibits hypersensitivity to calcofluor white (CW)¹⁰. An *MgAPT2* complementary DNA was expressed in the Δ *dnf1* Δ *dnf2* Δ *dnf3* triple mutant PFY3273A under the control of the *GAL1* promoter. Expression of *MgAPT2* in two independent yeast transformants restored the ability of PFY3273A to grow in the presence of 2.5 μ g ml⁻¹ CW (Fig. 1c). *MgAPT2* therefore encodes a P-type ATPase that is functionally related to the Drs2 family of APTs. However, *MgAPT2* failed to complement the temperature-sensitive growth phenotype of a *drs2* Δ mutant (Supplementary Fig. 1) and did not show plasma membrane aminophospholipid translocase activity on expression in yeast (Supplementary Fig. 2), indicating that in yeast *MgApt2* is able to fulfil some, but not all, of the functions of the Drs2 family of APTs^{10,11} and is unlikely to be localized in the plasma membrane.

To test the role of *MgAPT2* in rice blast disease, a one-step gene replacement was performed in which the 4.7-kilobase (kb) coding sequence of *MgAPT2* was removed and replaced with a 1.4-kb gene cassette conferring hygromycin resistance¹³ (Supplementary Fig. 3). Growth of six independent Δ *Mgapt2* mutants revealed no significant effect on either hyphal development or sporulation (data not shown). However, the ability of Δ *Mgapt2* mutants to cause rice blast disease in the susceptible rice cultivar CO-39 was significantly affected (Fig. 2a and Supplementary Fig. 3). The mean density of disease lesions on leaves inoculated with the wild-type *M. grisea* strain Guy11 was 34.33 ± 3.6 lesions per 5 cm leaf tip compared with 2.23 ± 1.5 in seedlings inoculated with the Δ *Mgapt2* mutant TM2105.6 ($P < 0.0005$). All the Δ *Mgapt2* mutants showed less epidermal invasion than seedlings inoculated with Guy11 and TM2105.1 ($P < 0.0001$) (Supplementary Fig. 3d). To test whether the pathogenicity defect was associated with appressorium function, rice leaves were abraded to remove the cuticle, and spore suspensions were applied. Under these conditions, *M. grisea* invades plant tissue without the need for an appressorium⁵. In rice seedlings inoculated with Guy11 and TM2105.1, disease symptoms appeared within 48 h in wounded leaves. In contrast, leaves inoculated with Δ *Mgapt2* mutant TM2105.6 did not exhibit symptoms until 72–96 h after inoculation (Fig. 2c). Plant tissue colonization was therefore delayed

¹School of Biosciences, University of Exeter, Washington Singer Laboratories, Perry Road, Exeter EX4 4QG, UK.

in leaves infected with $\Delta Mgapt2$, even in the absence of appressorium formation. Reduced fungal biomass and sporulation was observed from wounded leaves (not shown). The ability of $\Delta Mgapt2$ to infect roots, which does not require appressorium formation, was also completely impaired (Fig. 2b), although the root surface was colonized normally (data not shown). Ultrastructural analysis of appressoria revealed that $\Delta Mgapt2$ mutants formed abnormal penetration hyphae compared with Guy11 and did not completely penetrate plant epidermal cells (Fig. 2d–g), often arresting growth before invasion of the epidermal cell.

In yeast, the Drs2p APT and the Drs2p-related P-type ATPases Dnf1p and Dnf2p have been implicated in endocytosis. Drs2p is a Golgi-localized APT that has a role in late Golgi function¹¹; *drs2Δ* mutants are synthetically lethal with clathrin heavy chain (*chc1*) mutants and *arf1Δ* ADP-ribosylation factor mutants, and show

endocytotic defects, particularly in endosome–vacuole docking^{10,11}. Dnf1p and Dnf2p are both plasma membrane APTs, and null mutants show a cold-sensitive defect in receptor-mediated and bulk-phase endocytosis¹⁴. Furthermore, Drs2p is required for the formation of a specific subset of clathrin-coated secretory vesicles, indicating a role in exocytosis¹⁰. To determine the role of MgApt2p, we investigated the subcellular location of the protein. A V5 viral epitope tag sequence was incorporated into the 3' end of the *MgAPT2* gene. Expression of the resulting MgApt2–V5 fusion protein, after transformation of an *M. grisea* $\Delta Mgapt2$ mutant, complemented all mutant phenotypes and restored the ability to cause rice blast disease (Fig. 2a). For localization studies we generated an *M. grisea* strain expressing both the MgApt2–V5 fusion protein and a fusion protein of Ktr1 with green fluorescent protein (GFP). *MgKTR1* encodes a 1,2- α -mannosyltransferase (*M. grisea* genome database accession number MG08692, Broad Institute, USA), which is homologous to the Golgi-localized *S. cerevisiae* Ktr1 protein¹⁵. Immunofluorescence showed a punctate intracellular distribution pattern of the V5 epitope¹⁶ (Fig. 3a, b) co-localizing with MgKtr1–GFP fluorescence, which is consistent with localization to the Golgi. In filamentous fungi, discrete stacks of Golgi cisternae are rarely observed; instead, a disperse, vesicular Golgi apparatus is normally seen¹⁶. Immunogold electron microscopy confirmed the localization of MgApt2p to an internal vesicular compartment (Fig. 3c, d). We also performed an

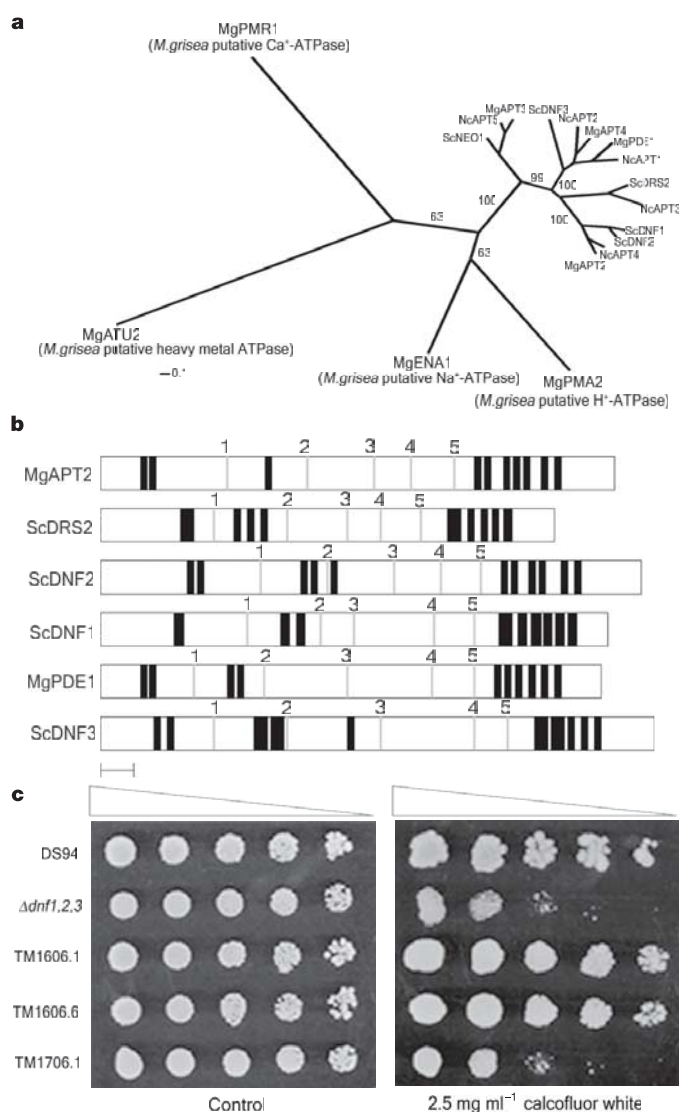


Figure 1 | Characteristics of *M. grisea* APT2. **a**, Phylogram of aminophospholipid translocases (APTs) from *M. grisea*, *S. cerevisiae* and *Neurospora crassa* (see Supplementary Information). **b**, Predicted topology of MgApt2, MgPde1 and yeast APTs. Thick lines indicate transmembrane helices. Thin lines are P-type ATPase motifs (1, LDGET; 2, DKTGLT phosphorylation site; 3, KGA nucleotide-binding site; 4, LTGD.ATP binding site; 5, GDGXND hinge motif). Scale bar, 100 amino acids. **c**, *MgAPT2* was expressed in the *S. cerevisiae* CW-sensitive $\Delta dnf1,2,3$ mutant (TM1606.1 and TM1606.6). TM1706.1 is $\Delta dnf1,2,3$ transformed with the empty vector. DS94 was used as wild type. The triangle indicates reduced concentration of the inoculum.

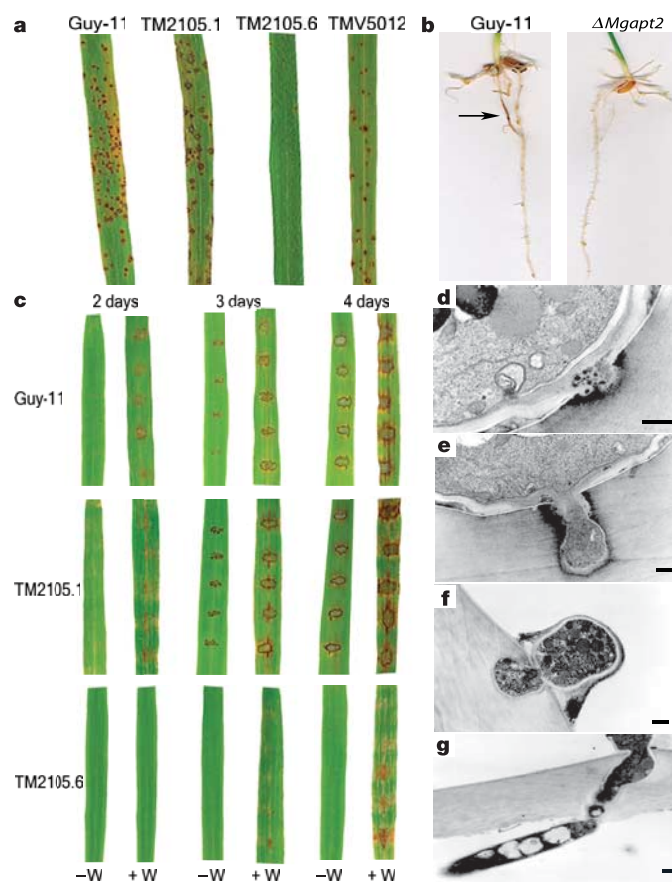


Figure 2 | Characterization of $\Delta Mgapt2$ mutants. **a**, Rice seedlings (cv. CO-39) were inoculated with Guy11, TM2105.1, TM2105.6 ($\Delta Mgapt2$) and TMV5012 ($\Delta Mgapt2$ expressing V5-epitope-tagged allele of *MgAPT2*). **b**, Blast symptoms on rice roots. Arrow shows necrotic lesion. **c**, Disease symptoms on intact (–W) and wounded leaf tissue (+W) inoculated with spore droplets. **d–g**, Transmission electron micrographs showing attempted penetration hypha development (**d**) and cuticle penetration (**e**) by TM2105.6 ($\Delta Mgapt2$) after 24 hours on onion epidermis compared with the same developmental stages in Guy11 (**f** and **g**). Scale bars, 500 nm.

ultrastructural examination of conidia, hyphae and appressoria of Δ Mgapt2 mutants after cryofixation, which revealed the presence of abnormal membrane-bound ring structures within the cytoplasm after cold shock (Fig. 3e, f). Direct comparison with a *drs2* Δ yeast mutant, previously subjected to cold shock, highlighted the similarity of these structures to Berkeley bodies¹¹ (Supplementary Fig. 4). To determine whether endocytosis was affected by the loss of MgApt2p, uptake of the lipophilic styryl dye FM4-64 was studied in Guy11 and Δ Mgapt2, but showed no difference between the strains (Supplementary Fig. 5).

However, we observed that Δ Mgapt2 mutants were unable to grow on a range of substrates when these were provided as sole carbon sources. Δ Mgapt2 mutants did not grow, for example, on starch, amylopectin or glycogen (Fig. 4a and Supplementary Table 2), and grew poorly on casein or lipids. We reasoned that these growth phenotypes could be a consequence of an exocytotic defect, preventing the secretion of extracellular depolymerizing enzymes. To test this idea we investigated the export of the starch-degrading enzyme α -amylase. Fungal mycelium was grown for 10 days in either glucose-containing medium or starch-containing medium, and α -amylase was detected in the culture filtrate with an enzyme-linked

immunosorbent assay (ELISA). The enzyme is induced only in the presence of starch and was readily detected at high concentration in culture filtrates of Guy11. In contrast, only trace amounts of α -amylase were detected in culture filtrates of the Δ Mgapt2 mutant (Fig. 4b, c). Immunogold electron microscopy with an anti- α -amylase antibody confirmed the secretion defect of Δ Mgapt2 mutants, and showed a predominantly cytoplasmic distribution of gold particles, which accumulated around the inside of the plasma membrane (Supplementary Fig. 6). However, only a subset of extracellular enzymes seem to require MgApt2p for secretion. An acidic trehalase, for instance, was secreted normally and the overall

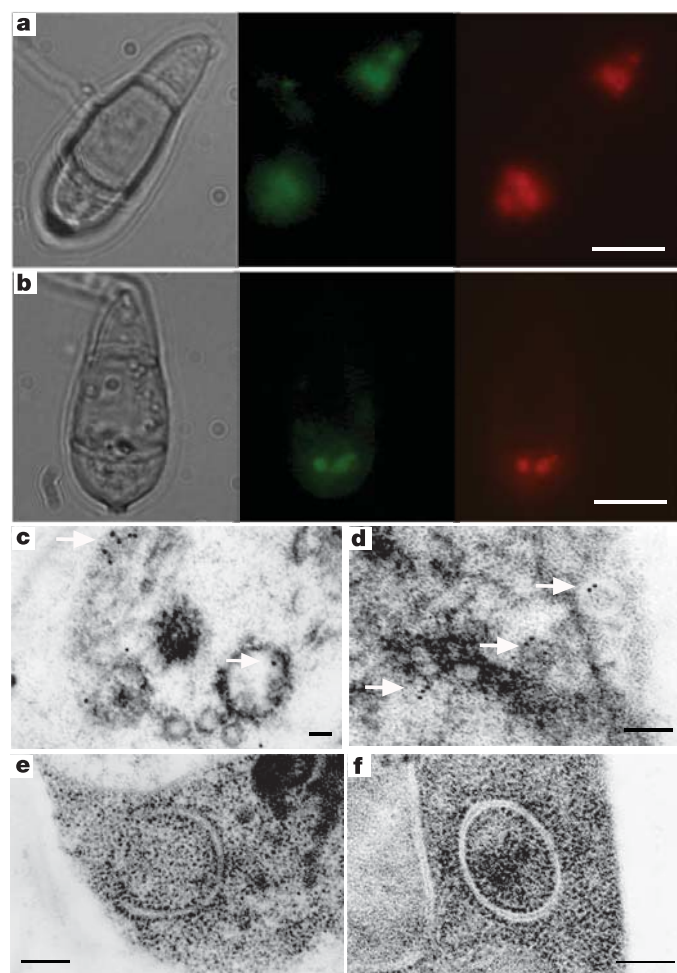


Figure 3 | Subcellular co-localization of MgApt2 and MgKtr1 Golgi protein. **a, b**, Micrographs of *M. grisea* conidium (left) expressing MgKtr1-GFP (middle) and MgApt2-V5 epitope-tagged allele (right) detected with Texas red (see Methods). Scale bars, 10 μ m. **c, d**, Representative immunogold electron micrographs of conidia from TMV5012 with the use of anti-V5 antibody. Scale bar, 100 nm (**c**); 250 nm (**d**). **e, f**, Transmission electron micrographs of TM2105.6 (Δ Mgapt2) hyphae, incubated at 4 °C for 4 h before fixation, showing typical accumulation of circular intracellular membrane structures. Scale bars, 150 nm.

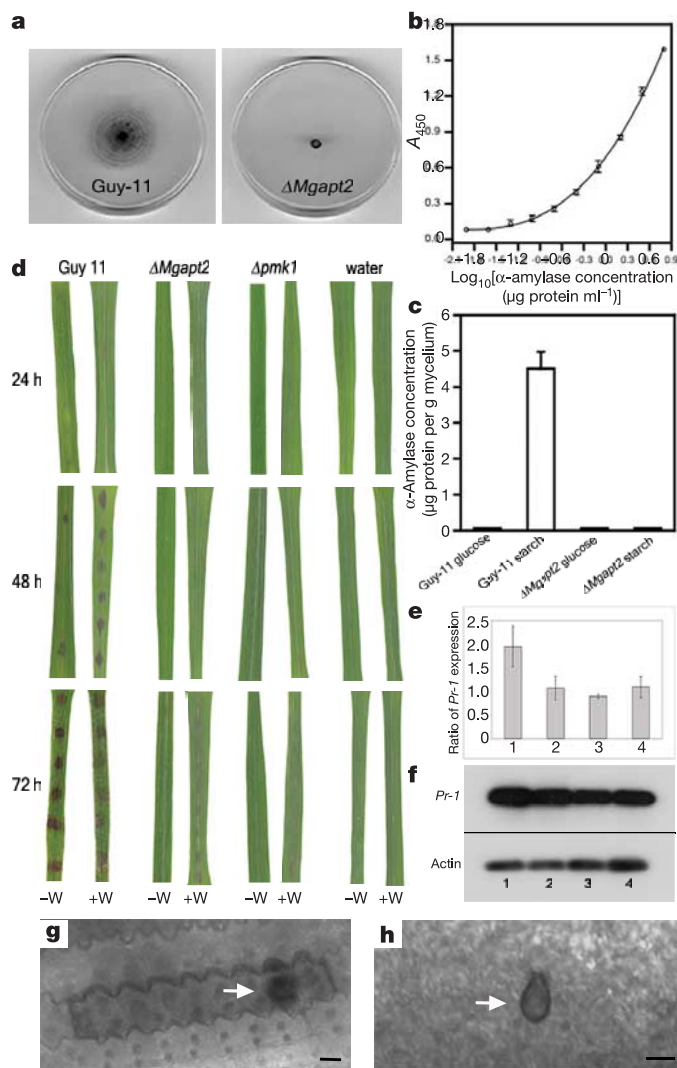


Figure 4 | Δ Mgapt2 mutants are impaired in secretion and induction of rice defence responses. **a**, Impaired growth of Δ Mgapt2 on starch-containing medium. **b, c**, Immunodetection of α -amylase in culture filtrates of Guy11 and Δ Mgapt2 grown in glucose or starch. Enzyme concentrations in **c** were determined by converting absorbances from ELISA with anti-fungal α -amylase serum to equivalents of protein concentration by using a calibration curve (**b**) of purified fungal α -amylase. **d**, Elicitation of hypersensitive reaction in rice cultivar IR-68 in non-wounded (–W) and wounded (+W) tissues. **e**, Bar chart showing the expression of rice *Pr-1* relative to actin gene expression 48 h after challenge by Guy11 (1), Δ Mgapt2 (2), Δ pmk1 (3) and water (4), based on RT-PCR analysis shown in (**f**). Total RNA and resulting cDNA were quantified with an Agilent 2100 Bioanalyser. Values in **b, c** and **e** are means $\pm$ s.d. for three repetitions of the experiment. **g, h**, Hypersensitive cell death in rice detected with Trypan blue, after challenge with Guy11 (**g**) or Δ Mgapt2 (**h**). Appressoria indicated by arrows. Scale bars, 10 μ m.

amount of protein secreted from growing hyphae was not affected by the loss of MgApt2p (Supplementary Fig. 7). To determine whether other virulence-associated proteins require MgApt2 for exocytosis, we characterized the behaviour of Δ Mgapt2 in an incompatible interaction with its host. Rice cultivars resistant to *M. grisea* often exhibit a hypersensitive reaction in response to challenge with the pathogen. Rice and *M. grisea* show a gene-for-gene interaction in which single major genes for resistance in the host are required for the recognition of pathogen-derived molecules encoded by fungal avirulence genes^{2,5}. In *M. grisea* several avirulence gene products are now known⁸ and in AvrPi-ta and its cognate resistance protein Pi-ta a direct interaction within rice cells has been shown¹⁷. We inoculated rice cultivar IR-68, which is resistant to *M. grisea* Guy11, with a Δ Mgapt2 mutant and found an inability to elicit the hypersensitive reaction after 48 h (Fig. 4d). We reasoned that this might be due to the inefficiency of appressorium-mediated cuticle penetration of Δ Mgapt2 mutants. We therefore removed the cuticle by abrasion and inoculated IR-68 leaves with Guy11, a Δ Mgapt2 mutant and a non-appressorium-forming Δ pmk1 mutant¹⁸. Even after abrasion, hypersensitivity did not become apparent until 72 h after inoculation with a Δ Mgapt2 mutant (Fig. 4d). To investigate the plant defence response, the expression of the rice defence gene *Pr-1* (ref. 19) was analysed in IR-68 tissue by semiquantitative polymerase chain reaction with reverse transcription (RT-PCR) and found to be significantly decreased ($P < 0.05$) on challenge with the Δ Mgapt2 mutant compared with Guy11 (Fig. 4e, f). At the cellular level, hypersensitive cell death in IR68 as a response to Guy11 was observed by Trypan blue staining and occurred at a much lower frequency in response to Δ Mgapt2 infections (Fig. 4g, h). When considered together, these results indicate that the secretion of fungal proteins perceived by the rice plant host during a resistance response might also require MgApt2.

Characterization of the MgApt2 P-type ATPase has consequences for our understanding of fungal pathogenicity and the development of rice blast disease. It is clear from the genome sequence of *M. grisea* that a family of APTs exists in the fungus. Previous work has established that Pde1p (a functional homologue of yeast Dnf3p) acts as a virulence determinant in *M. grisea* by affecting appressorium function²⁰. The current report has shown that MgApt2, an APT with a role in exocytosis, is required for plant tissue colonization. Taken together, the studies provide evidence that APTs regulate membrane characteristics important to fungal pathogens, both for infection-related morphogenesis and for the efficient delivery of virulence-associated proteins. The identification of MgApt2p will allow the dissection of the processes required for secretion by pathogenic fungi during plant infection and will provide a method of identifying effector proteins essential for rice blast disease.

METHODS

Full details are given in Supplementary Methods.

Fungal growth conditions. All *M. grisea* strains used in this study are derived from Guy11 and were maintained as described previously. Yeast strains were manipulated in accordance with standard methods²¹.

Expression of MgApt2 in *S. cerevisiae*. A full-length MgAPT2 cDNA was obtained by rapid amplification of cDNA ends (RACE)²² with a Clontech SMART kit. Individual 3' and 5' RACE products were cloned into the pGEM-T vector (Promega), and the full-length cDNA was sequenced. DNA sequencing used a Global IR2 NEN automated DNA sequencer (LI-COR) and fluorescent infrared-labelled primers. The 4.5-kb MgAPT2 cDNA was cloned into pYES-2 (Invitrogen) and introduced into the *S. cerevisiae* dnf1,2,3 Δ mutant²¹. Transformants were confirmed by PCR. Sensitivity to CW¹¹ was assessed by plating a dilution series of yeast cells (10^6 – 10^4 cells ml⁻¹) on yeast extract/peptone (YP) plates⁹ containing 2.5 μ g ml⁻¹ CW in the presence of galactose for 17 h at 29 °C.

Targeted gene replacement of MgAPT2. An MgAPT2 gene replacement vector was constructed by amplifying 1.3-kb and 1.8-kb flanking sequences of the MgAPT2 coding sequence by PCR from Guy11 genomic DNA with 35 cycles. The products were cloned into pGEM-T and a BamHI-linked hygromycin

phosphotransferase selectable marker gene¹³ ligated between the two flanking sequences⁸. The linear 4.5-kb gene replacement construct was excised with *ApaI* and *SpeI* and introduced into *M. grisea*²³. Putative transformants were confirmed by Southern blots, as described previously.

Rice infections. Rice infections were performed with a susceptible Indica rice (*Oryza sativa*) cultivar CO-39 (ref. 24) or the resistant cultivar IR-68 (ref. 25). Uniform conidial suspensions were prepared and applied to rice leaves as described previously²³. Appressorium-mediated penetration was assessed on onion (*Allium cepa*) epidermal strips²⁶. Infection of wounded rice leaf tissue was performed by gently abrading the leaf surface with glasspaper before placing leaf sections on water agar plates (4% w/v) and inoculating with 10 μ l droplets containing 10^4 conidia ml⁻¹.

Subcellular co-localization of MgDRS2p and MgKTR-GFP. A 6.6-kb genomic DNA fragment spanning the MgAPT2 protein-coding region and 1.6-kb upstream sequence was amplified and cloned into pTOPO 2.1 (Invitrogen). A V5 epitope was incorporated by performing inverse PCR and ligation of the resulting amplicon¹⁶. The successful incorporation of the V5 epitope was confirmed by DNA sequencing. A sulphonylurea resistance gene²⁷ was cloned into the BamHI site of pMJGgAPT5 and the plasmid was transformed into the Δ Mgapt2 mutant. Sulphonylurea-resistant transformants were selected and analysed by Southern blotting. Co-localization studies were performed by generating a MgKtr1–GFP gene fusion. A 3.5-kb fragment of the *MgKTR1* gene was isolated and an *XhoI* site was engineered at the 3' end of the gene to allow an in-frame fusion to the sGFP gene (see Supplementary Methods for details). The resulting construct was transformed into an *M. grisea* strain expressing pMJGgAPT5.

Immunological procedures. Immunofluorescence microscopy was performed as described²⁸. Conidia (10^4 ml⁻¹) were allowed to germinate and form appressoria on multi-test ten-well glass slides (ICN) for 2–18 h. To allow antibodies access to internal structures, enzymatic digestion of the cell wall with Glucanex (50 mg ml⁻¹) was performed for 90 min at 20 °C. Anti-V5 antibody (Invitrogen) and Texas red-conjugated anti-mouse secondary antibody were added as described in Supplementary Methods. Samples were viewed on a Zeiss Axioskop 2 microscope with the Axiovision 3.0 software. For immunogold electron microscopy, procedures were as described previously²⁸.

Received 23 September; accepted 22 December 2005.

- Alfano, J. R. & Collmer, A. Type III secretion system effector proteins: Double agents in bacterial disease and plant defense. *Annu. Rev. Phytopathol.* **42**, 385–414 (2004).
- Nimchuk, Z., Eulgem, T., Holt, B. E. & Dangl, J. L. Recognition and response in the plant immune system. *Annu. Rev. Genet.* **37**, 579–609 (2003).
- Ghosh, P. Process of protein transport by the type III secretion system. *Microbiol. Molec. Biol. Rev.* **68**, 771–795 (2004).
- Mendgen, K., Hahn, M. & Deising, H. Morphogenesis and mechanisms of penetration by plant pathogenic fungi. *Annu. Rev. Phytopathol.* **34**, 367–386 (1996).
- Talbot, N. J. On the trail of a cereal killer: investigating the biology of *Magnaporthe grisea*. *Annu. Rev. Microbiol.* **57**, 177–202 (2003).
- Howard, R. J., Ferrari, M. A., Roach, D. H. & Money, N. P. Penetration of hard substrates by a fungus employing enormous turgor pressures. *Proc. Natl Acad. Sci. USA* **88**, 11281–11284 (1991).
- Sesma, A. & Osbourn, A. E. The rice leaf blast pathogen undergoes developmental processes typical of root-infecting fungi. *Nature* **431**, 582–586 (2004).
- Dean, R. A. et al. The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* **434**, 980–986 (2005).
- Chen, C.-Y., Ingram, M. F., Rosal, P. H. & Graham, T. R. Role for Drs2p, a P-type ATPase and potential aminophospholipid translocase in yeast late Golgi function. *J. Cell Biol.* **147**, 1223–1236 (1999).
- Gall, W. E. et al. Drs2p-dependent formation of exocytotic clathrin-coated vesicles in vivo. *Curr. Biol.* **12**, 1623–1627 (2002).
- Hua, Z., Fatheuddin, P. & Graham, T. R. An essential subfamily of Drs2-related P-type ATPases is required for protein trafficking between Golgi complex and endosomal/vacuolar system. *Mol. Biol. Cell* **13**, 3162–3177 (2002).
- Lutsenko, S. & Kaplan, J. H. Organization of P-type ATPases: Significance of structural diversity. *Biochemistry* **34**, 15607–15613 (1995).
- Carroll, A. M., Sweigard, J. A. & Valent, B. Improved vectors for selecting resistance to hygromycin. *Fungal Genet. Newsl.* **42**, 22 (1994).
- Pomorski, T. et al. Drs2-related P-type ATPases Dnf1p and Dnf2p are required for phospholipids translocation across the yeast plasma membrane and serve a role in endocytosis. *Mol. Biol. Cell* **14**, 1240–1254 (2003).
- Lussier, M., Sdicu, A. M., Camirand, A. & Bussey, H. Functional Characterization of the YUR1, KTR1, and KTR2 genes as members of the yeast KRE2/MNT1 mannosyltransferase gene family. *J. Biol. Chem.* **271**, 11001–11008 (1996).
- Southern, J. A., Young, D. F., Heaney, F., Baumgartner, W. & Randall, R. E.

- Identification of an epitope on the P and V proteins of simian virus 5 that distinguishes between two isolates with different biological characteristics. *J. Gen. Virol.* **72**, 1551–1557 (1991).
17. Jia, Y., McAdams, S. A., Bryan, G. T., Hershey, H. P. & Valent, B. Direct interaction of resistance gene and avirulence gene products confers rice blast resistance. *EMBO J.* **19**, 4004–4014 (2000).
 18. Xu, J. R. & Hamer, J. E. MAP kinase and cAMP signalling regulate infection structure formation and pathogenic growth in the rice blast fungus *Magnaporthe grisea*. *Genes Dev.* **10**, 2696–2706 (1996).
 19. Kim, S. *et al.* Molecular characterization of the cDNA encoding an acidic isoform of PR-1 protein in rice. *Mol. Cells* **11**, 115–121 (2001).
 20. Balhadère, P. V. & Talbot, N. J. Pde1 encodes a P-type ATPase involved in appressorium-mediate plant infection by the rice blast fungus *Magnaporthe grisea*. *Plant Cell* **13**, 1987–2004 (2001).
 21. Rose, M. D., Winston, F. & Hieter, P. *Methods in Yeast Genetics: A Laboratory Course Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1990).
 22. Frohman, M. A., Dush, M. K. & Martin, G. R. Rapid production of full-length cDNA from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proc. Natl Acad. Sci. USA* **85**, 8998–9002 (1988).
 23. Talbot, N. J., Ebbold, D. J. & Hamer, J. E. Identification and characterization of *MPG1*, a gene involved in pathogenicity from the rice blast fungus *Magnaporthe grisea*. *Plant Cell* **5**, 1575–1590 (1993).
 24. Valent, B., Farrall, L. & Chumley, F. G. *Magnaporthe grisea* genes for pathogenicity and virulence identified through a series of backcrosses. *Genetics* **127**, 87–101 (1991).
 25. Zeng, L. R. *et al.* Spotted leaf11, a negative regulator of plant cell death and defense, encodes a U-box/armadillo repeat protein endowed with E3 ubiquitin ligase activity. *Plant Cell* **17**, 2795–2808 (2004).
 26. Chida, T. & Sisler, H. D. Restoration of appressorial penetration ability by melanin precursors in *Pyricularia oryzae* treated with antipenetrants and in melanin-deficient mutants. *J. Pestic. Sci.* **12**, 49–55 (1987).
 27. Sweigard, J. A., Chumley, F. G., Carroll, A., Farrall, L. & Valent, B. A series of vectors for fungal transformation. *Fungal Genet. Newsl.* **44**, 52–53 (1997).
 28. Tucker, S. L. *et al.* A fungal metallothionein is required for pathogenicity of *Magnaporthe grisea*. *Plant Cell* **16**, 1575–1588 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Hawes and B. Martin for cryofixation and freeze substitution work, G.-L. Wang for supply of IR-68 seeds and T. Graham for supplying the BY4739, PFY3273A and DS94 yeast APT mutants. This study was supported by a grant to N.J.T. from the Biological Sciences and Biotechnology Research Council (BBSRC).

Author Contributions Experimental work and data analysis were performed by M.J.G. and N.J.T. C.R.T. performed all immunological work and associated data analysis. G.E.W. performed electron microscopy.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to N.J.T. (n.j.talbot@exeter.ac.uk).

LETTERS

CD69 acts downstream of interferon- α/β to inhibit S1P₁ and lymphocyte egress from lymphoid organs

Lawrence R. Shioh^{1,2,4}, David B. Rosen^{2,4}, Naděžda Brdičková³, Ying Xu^{1,2}, Jinping An^{1,2}, Lewis L. Lanier², Jason G. Cyster^{1,2} & Mehrdad Matloubian³

Naive lymphocytes continually enter and exit lymphoid organs in a recirculation process that is essential for immune surveillance. During immune responses, the egress process can be shut down transiently¹. When this occurs locally it increases lymphocyte numbers in the responding lymphoid organ; when it occurs systemically it can lead to immunosuppression as a result of the depletion of recirculating lymphocytes. Several mediators of the innate immune system are known to cause shutdown, including interferon α/β (IFN- α/β) and tumour necrosis factor^{2–5}, but the mechanism has been unclear. Here we show that treatment with the IFN- α/β inducer polyinosine polycytidylic acid (hereafter ‘poly(I:C)’) inhibited egress by a mechanism that was partly lymphocyte-intrinsic. The transmembrane C-type lectin CD69 was rapidly induced and CD69^{−/−} cells were poorly retained in lymphoid tissues after treatment with poly(I:C) or infection with lymphocytic choriomeningitis virus. Lymphocyte egress requires sphingosine 1-phosphate receptor-1 (S1P₁), and IFN- α/β was found to inhibit lymphocyte responsiveness to S1P. By contrast, CD69^{−/−} cells retained S1P₁ function after exposure to IFN- α/β . In coexpression experiments, CD69 inhibited S1P₁ chemotactic function and led to downmodulation of S1P₁. In a reporter assay, S1P₁ crosslinking led to co-crosslinking and activation of a CD69–CD3 ζ chimera. CD69 co-immunoprecipitated with S1P₁ but not the related receptor, S1P₃. These observations indicate that CD69 forms a complex with and negatively regulates S1P₁ and that it functions downstream of IFN- α/β , and possibly other activating stimuli, to promote lymphocyte retention in lymphoid organs.

When mice were treated with poly(I:C), a double-stranded RNA mimetic, we observed that lymphocyte numbers in blood and lymph decreased rapidly, reaching a minimum by 6 h (Fig. 1a, and Supplementary Fig. 1a). This effect was largely mediated by IFN- α/β because mice deficient in IFN- α/β -receptor-1 (IFNAR1) showed only a weak response (Fig. 1b). The effect of IFN- α/β was, in part, lymphocyte-intrinsic because adoptively transferred *Ifnar1*^{−/−} lymphocytes were less efficiently sequestered than wild-type cells after treatment with poly(I:C) (Fig. 1c, and Supplementary Fig. 1b). In particular, *Ifnar1*^{−/−} T cells egressed into the lymph of poly(I:C)-treated animals in fourfold greater numbers than wild-type control cells present in the same animals, and *Ifnar1*^{−/−} B cells entered the lymph in sixfold to sevenfold greater numbers (Fig. 1c, d).

The recent finding that S1P₁, a G-protein-coupled receptor (GPCR), is required within lymphocytes for egress from lymphoid organs^{6,7} led us to ask whether IFNAR1 signalling altered lymphocyte S1P responsiveness. In chemotaxis assays, T lymphocytes taken from mice that had been pretreated with poly(I:C) showed markedly reduced responsiveness to S1P (Fig. 1e). Lymphocytes that lacked IFNAR1 were largely resistant to this effect, indicating that it was due to intrinsic IFNAR1 signalling (Fig. 1e). Although B cells migrate

poorly to S1P directly *ex vivo*, they acquire increased chemotactic responsiveness after several hours *in vitro* incubation (Fig. 1f). Exposure of the splenocyte cultures to poly(I:C) or to recombinant IFN- α/β inhibited the B-cell response to S1P (Fig. 1f). In contrast with the marked loss of S1P chemotactic responsiveness, quantitative polymerase chain reaction (PCR) analysis showed that S1P₁ mRNA was only partly decreased in abundance in T and B cells from poly(I:C)-treated mice (Fig. 1g).

IFN- α/β is a strong inducer of the cell surface activation marker, CD69 (Fig. 2a). In two independent studies, it was found that transgenic overexpression of CD69 causes decreased egress of mature single-positive T cells from the thymus^{8,9}. However, CD69-deficient mice had normal thymic egress, calling into question the physiological relevance of the transgenic studies^{10,11}. We tested whether CD69 might be involved in the IFNAR1-mediated inhibition of lymphocyte egress by transferring wild-type and CD69^{−/−} cells into wild-type hosts and treating them with poly(I:C). Whereas transferred wild-type cells were almost undetectable in the lymph 6 h after treatment with poly(I:C), CD69^{−/−} cells continued to appear in the lymph (Fig. 2b). Similarly, the sequestration and depletion of T and B lymphocytes from blood was less complete for CD69^{−/−} cells than for wild-type cells (Supplementary Fig. 2a). Quantitative PCR analysis for interferon response factor (IRF)-7 confirmed that the CD69^{−/−} and wild-type cells were similarly activated (Supplementary Fig. 2b) and S1P₁ mRNA abundance was reduced to the same extent as that observed in wild-type cells (Supplementary Fig. 2c). These observations showed that CD69 is required downstream of IFNAR1 to inhibit lymphocyte egress. Lymphocytic choriomeningitis virus (LCMV) infection also caused a decrease in T-cell and B-cell numbers in lymph and blood, and this effect was strongly inhibited when the lymphocytes lacked CD69 (Fig. 2c, and Supplementary Fig. 3a, b).

In contrast to wild-type T cells, CD69^{−/−} T cells retained a substantial fraction of their S1P responsiveness after treatment with poly(I:C) (Fig. 2d). Similarly, CD69^{−/−} B cells exposed to IFN- α/β *in vitro* retained much of their S1P responsiveness (Fig. 2d). By flow cytometric analysis, S1P₁ was downregulated from the surface of wild-type T cells but was only partly downregulated on CD69^{−/−} T cells after treatment with poly(I:C) (Fig. 2e) or infection with LCMV (Supplementary Fig. 3c). These observations indicated that CD69 might be required for maximal S1P₁ downmodulation.

To further characterize the mechanism of CD69-mediated inhibition of S1P₁, we used a co-transfection approach in the WEHI-231 B-cell line. Cells were transduced with a retrovirus encoding Flag-tagged mouse S1P₁, an internal ribosomal entry site (IRES) and cytoplasmic-domain truncated human CD4 (hCD4) as a reporter. These cells expressed the Flag-tagged receptor on the surface in amounts proportional to the amount of hCD4 expressed (Fig. 3a).

¹Howard Hughes Medical Institute, ²Department of Microbiology and Immunology, ³Department of Medicine, and ⁴Biomedical Sciences Graduate Program, University of California, San Francisco, San Francisco, California 94143, USA.

Similarly, cells transduced with a mouse CD69-IRES- green fluorescent protein (GFP) construct expressed CD69 on the surface in proportion to the amount of GFP expressed (Fig. 4a). However, when the CD69-IRES-GFP retrovirus was introduced into the Flag-S1P₁-expressing cells, there was a loss of surface Flag-S1P₁ staining in cells expressing the highest amounts of GFP (Fig. 3a, b), and in chemotaxis assays these cells showed little response to S1P but retained responsiveness to the chemokine SDF1 (CXCL12) (Fig. 3c). Co-transduction of Flag-S1P₁-expressing cells with another C-type lectin, mouse CD23, did not lead to downmodulation of surface expression and did not affect the function of S1P₁ (Fig. 3b, c). When CD69 was introduced into cells expressing Flag-S1P₃, the receptor was not downmodulated (Fig. 3b). These findings indicate that in cells with abundant CD69, S1P₁ is selectively downmodulated and functionally inhibited.

Although CD69 has been well studied as a lymphocyte activation marker and is transcriptionally induced by many activation stimuli^{12,13}, naive CD69-surface negative lymphocytes express considerable amounts of basal CD69 mRNA (Figs 2a and 3d) and have detectable intracellular CD69 protein¹⁴. In S1P₁-deficient B and T cells, CD69 is found on the cell surface, although these cells do not express other activation markers or show evidence of increased cell size^{6,15}. Moreover, S1P₁-deficient B and T cells have the same amounts of CD69 mRNA as control B and T cells (Fig. 3d). These findings indicate that in naive wild-type lymphocytes the small amounts of CD69 that are expressed might be prevented from surface expression by S1P₁. Indeed, when we plotted Flag-S1P₁ against CD69 in total GFP⁺ hCD4⁺ co-transduced WEHI-231 cells we observed that there was mutual exclusion of surface expression (Fig. 3e). The CD69⁺ Flag-S1P₁⁻ cells (Fig. 3e, region 1) expressed high levels of

GFP (Fig. 3a, b); by contrast, cells expressing only Flag-S1P₁ on the surface were GFP^{low} (Fig. 3e, region 2 and inset). CD69 was readily detectable on the surface of GFP^{low} cells after transduction with the CD69 construct alone (Fig. 3f). Thus, in addition to the inhibitory effect of high CD69 expression on surface Flag-S1P₁ levels, there was an inhibitory effect of S1P₁ on surface CD69 expression when the abundance of CD69 was low. This effect was specific for CD69 and S1P₁ because CD69 and S1P₃ were coexpressed on the surface of cells without any evidence of cross-inhibition (Fig. 3g). In concordance with these observations, human S1P₁ was identified in a retroviral cDNA library overexpression screen as a suppressor of human CD69 surface expression¹⁶. S1P₁ is a G_i-coupled receptor¹⁷, and treatment of the cells co-transduced with S1P₁ and CD69 for 6 h with pertussis toxin, a G_i inhibitor, led to partial recovery of CD69 surface expression in the S1P₁⁺ GFP^{low} cells (Fig. 3e, h). Thus, S1P₁-mediated modulation of CD69 was at least partly dependent on G_i coupling. Agonist-induced downmodulation might account for the ability of the S1P receptor agonist FTY720-phosphate to cause CD69 downmodulation in single-positive thymocytes¹⁸ that normally reside in an S1P-poor environment¹⁹.

To test further for an association between CD69 and S1P₁, 2B4 T lymphoma cells expressing a nuclear factor of activated T cells (NFAT)-GFP reporter²⁰ were transduced with a mouse CD69 ectodomain and transmembrane domain plus human CD3 ζ cytoplasmic domain fusion protein (CD69-CD3 ζ) alone or together with Flag-tagged S1P₁. In this assay, GFP expression is induced after CD3 ζ cytoplasmic domain cross-linking and, as expected, GFP induction was observed after treatment of the CD69-CD3 ζ -transduced cells with anti-CD69 antibodies (Fig. 4a). GFP induction was also observed after anti-Flag antibody treatment of cells co-transduced

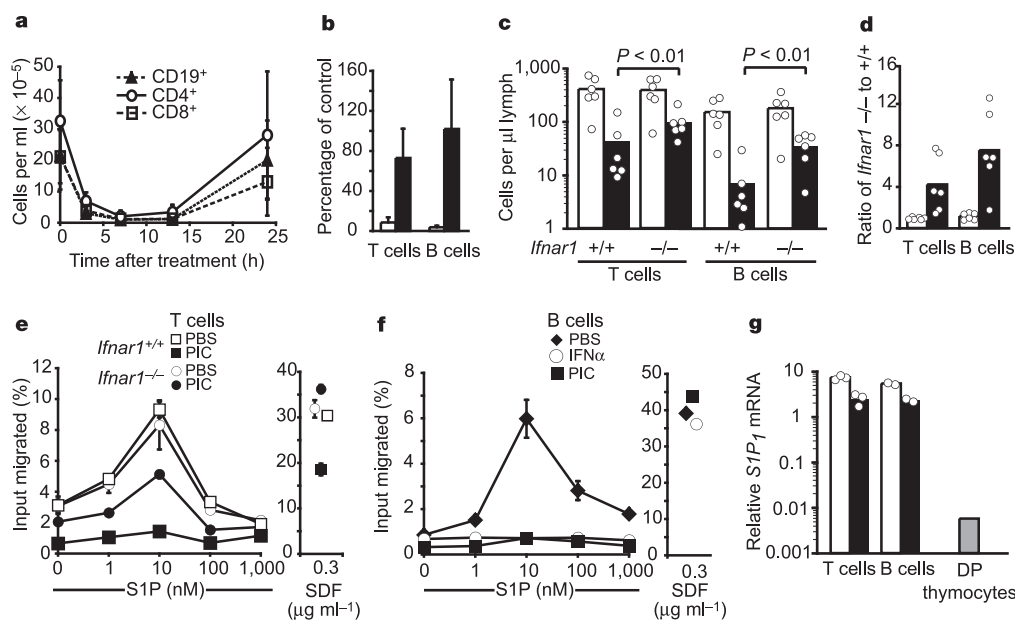


Figure 1 | Lymphocyte-intrinsic IFNAR1 requirement during poly(I:C)-induced lymphocyte sequestration and downregulation of S1P₁ by IFN- α/β . **a**, Mice were injected with phosphate-buffered saline (PBS) or poly(I:C) and at the indicated times B-cell and T-cell numbers in the lymph were determined. Results show means $\pm$ s.d. for three poly(I:C)-treated mice per time point. The '0' h data are from mice injected 6 h earlier with PBS. **b**, Wild-type mice (open columns) and *Ifnar1*^{-/-} mice (filled columns) were treated with poly(I:C) for 6 h; cell frequency in lymph is shown as a percentage of the PBS-treated controls ($n = 4$). **c**, **d**, Wild-type and IFNAR1-deficient lymphocytes were labelled with CFSE and CMTMR, respectively, and transferred into wild-type mice. At 6 h after treatment with poly(I:C) (filled columns) or PBS (open columns), lymph was analysed

for the number of transferred cells (**c**) and the ratio of transferred IFNAR1-deficient cells to wild-type cells (**d**). Columns represent averages and open circles individual animals from two experiments, representative of five. **e**, Migration assay of transferred T cells from recipients treated as in **c**. **f**, Migration assay of B cells from wild-type mice after incubation for 10 h in vehicle (diamonds), 20 $\mu\text{g ml}^{-1}$ poly(I:C) or 1,000 U ml^{-1} IFN- α . In **e** and **f**, data points show the mean of duplicates and bars the range for one of three experiments. **g**, Quantitative PCR analysis of S1P₁ in T and B cells from wild-type mice treated with poly(I:C) (filled columns) or PBS (open columns). Double-positive thymocytes from an untreated mouse are shown (grey column) as an example of cells lacking S1P₁ (ref. 6).

with CD69–CD3 ζ and Flag–S1P₁, whereas no GFP induction occurred after anti-Flag treatment of cells transduced with Flag–S1P₁ alone (Fig. 4a). In cells transduced with Flag–S1P₃ and CD69–CD3 ζ , no GFP induction occurred after treatment with anti-Flag antibody (Fig. 4a). Flow cytometric analysis demonstrated that the CD69–CD3 ζ chimera was coexpressed on the cell surface together with Flag–S1P₁ (Supplementary Fig. 4a). This was also observed when WEHI231 cells were co-transduced with these constructs (Supplementary Fig. 4b), indicating that the intact CD69 cytoplasmic domain is needed for efficient downmodulation of the coexpressed molecules from the cell surface. These findings indicate that antibody-mediated crosslinking of Flag–S1P₁ led to co-crosslinking and activation of CD69–CD3 ζ , providing further evidence that these molecules closely associate on the membrane and indicating that the CD69 ecto-domain and transmembrane domains are sufficient to mediate this association.

As a further approach to test whether S1P₁ and CD69 formed a complex, we examined whether these molecules could be coimmunoprecipitated. Using lysates prepared from co-transduced WEHI231 B cells we found that mouse CD69 co-precipitated with Flag–S1P₁, whereas another C-type lectin, CD23, did not (Fig. 4b). CD69 failed to co-immunoprecipitate with Flag–S1P₃ (Fig. 4b). Mouse CD69, CD23, S1P₁ and S1P₃ have N-linked glycosylation sites, and the broad band detected for each molecule in western blot analysis is most probably due to the glycosylation. To test whether an association also occurred between human CD69 and S1P₁, we prepared lysates from Jurkat T lymphoma cells that had been transduced with epitope-tagged human S1P₁ and stimulated with anti-T cell receptor antibody to promote the upregulation of CD69. Immunoprecipitation of endogenous CD69 was found to co-immunoprecipitate

epitope-tagged S1P₁ and vice versa (Fig. 4c). Although not excluding the involvement of additional molecules, these findings provide evidence for an evolutionarily conserved interaction between CD69 and S1P₁ in lymphocytes.

Our observations indicate that CD69, a type II transmembrane protein, might engage in a novel mode of GPCR negative regulation by forming a complex with S1P₁, promoting downmodulation and inhibiting the function of this receptor in lymphocyte egress. In addition to IFN- α/β , CD69 transcription is induced in response to most stimuli that cause lymphocyte activation, including other stimuli that are known to cause lymphocyte sequestration, such as tumour necrosis factor and antigen stimulation^{5,12,21}. Activating stimuli can also cause S1P₁ transcriptional downregulation^{6,22} and this would be expected to decrease the amounts of CD69 needed to dominate and inhibit any remaining S1P₁. Consistent with CD69 functioning to promote sequestration of human lymphocytes is the observation that human CD69⁺ T cells are rare in blood compared with lymphoid tissues and inflammatory sites¹². In addition to lymphocyte-intrinsic effects, our findings indicate that IFN- α/β has extrinsic effects that regulate egress. CD69 deficiency has effects on anti-tumour responses and the course of collagen-induced arthritis¹³. This has been suggested to be due to altered production of transforming growth factor- β ²³, but altered trafficking of activated lymphocytes may also be involved. We speculate that in some cell types, S1P₁ and CD69 may be coexpressed as a non-modulating complex and that this might allow CD69 signalling through associated S1P₁ molecules. In this regard, it is notable that CD69 has been found to be associated with a heterotrimeric G protein¹⁴ and that some effects of CD69 cross-linking antibodies are sensitive to pertussis toxin^{24–26}. S1P₁ is currently being targeted by small

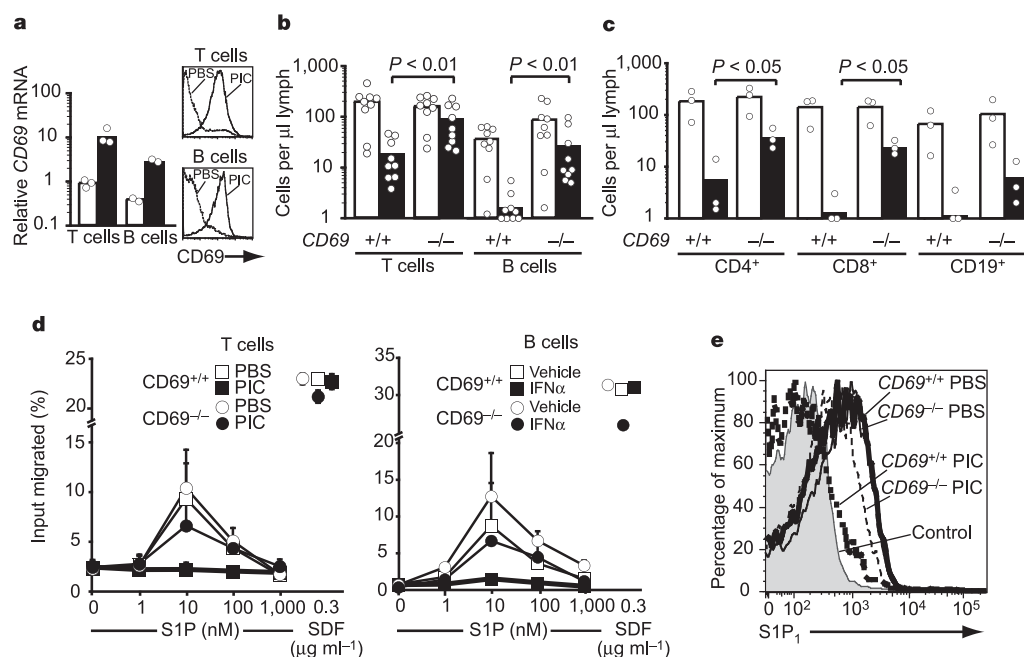


Figure 2 | CD69-deficient cells are less efficiently sequestered after treatment with poly(I:C) and LCMV infection and retain S1P responsiveness. **a**, Quantitative PCR (left panel) and flow cytometric analysis (right panels) of CD69 in spleen T and B cells from mice treated with poly(I:C) (filled columns) or PBS (open columns). Staining of cells from PBS-treated mice was identical to the staining of CD69^{-/-} cells (not shown). PIC, poly(I:C). **b**, Wild-type and CD69-deficient lymphocytes were labelled and transferred into recipient mice that were treated and analysed as in Fig. 1c. Filled columns, poly(I:C); open columns, PBS. **c**, Recipients of labelled wild-type and CD69-deficient lymphocytes were uninfected (open columns) or infected (filled columns) with LCMV and analysed at 18 h for

cell numbers in lymph. Data are representative of three experiments. **d**, Left: migration assay of transferred T cells from recipients of wild-type (squares) and CD69^{-/-} (circles) lymphocytes treated with PBS or poly(I:C). Data are means $\pm$ s.d. for nine mice from three experiments. Right: migration assay of B cells from wild-type (squares) and CD69^{-/-} (circles) mice after incubation for 6 h in the absence or presence of 1,000 U ml⁻¹ IFN- α . Data are from one experiment, representative of three. **e**, Flow cytometric analysis of S1P₁ on splenic CD4⁺ L-selectin^{high} T cells from the indicated mice after treatment with poly(I:C) (PIC) or PBS. Control antibody staining is shown shaded in grey. Data are representative of ten mice of each type analysed in six experiments.

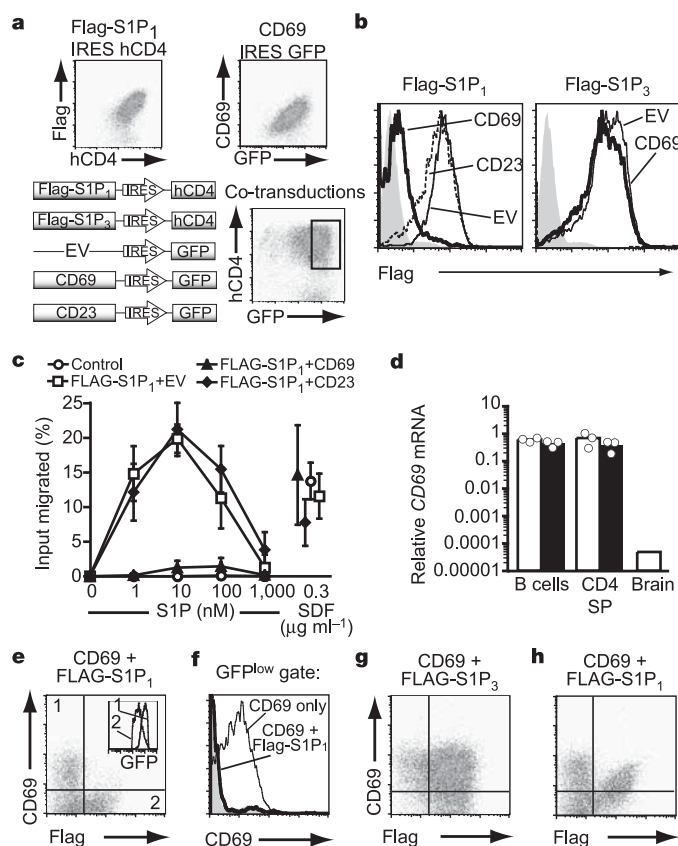


Figure 3 | CD69 interacts with S1P₁ causing downregulation and inhibition of S1P₁ function. **a**, Diagram of retroviral constructs, and flow cytometric analysis of WEHI-231 cells transduced and stained to detect Flag-S1P₁, CD69 or hCD4, or showing GFP fluorescence as indicated. The plot of co-transduced cells shows the GFP^{high} gating strategy used in **b** and **c**. **b**, Flag-S1P₁ or Flag-S1P₃ expression in GFP^{high} hCD4⁺ cells co-transduced with empty vector (EV), CD69 or CD23. The grey histogram shows staining of non-transduced WEHI-231 cells. Data are representative of three to six experiments. **c**, Migration assay of untransduced WEHI-231 cells (control) and GFP^{high} hCD4⁺ Flag-S1P₁ WEHI-231 cells co-transduced with EV, CD69 or CD23. Results are means ± s.d. for three experiments. **d**, Quantitative PCR analysis of CD69 abundance in sorted B cells and mature CD4 single-positive (SP) L-selectin^{high} thymocytes from mice reconstituted with wild-type (open columns) or S1P₁^{-/-} (filled columns) fetal liver cells. Transcripts obtained from total brain are shown for comparison. **e**, Flag-S1P₁ and CD69 expression on total GFP⁺ hCD4⁺ Flag-S1P₁ WEHI-231 cells co-transduced with CD69. Inset: GFP levels of cells in regions 1 and 2. **f**, CD69 expression on GFP^{low} cells transduced with CD69 alone or co-transduced with Flag-S1P₁. The grey histogram shows staining of non-transduced WEHI-231 cells. **g**, Flag-S1P₃ and CD69 expression on Flag-S1P₃ WEHI-231 cells co-transduced with CD69. **h**, Flag-S1P₁ and CD69 expression on Flag-S1P₁ WEHI-231 cells co-transduced with CD69 and incubated for 6 h with 200 ng ml⁻¹ pertussis toxin.

molecules as a mechanism of inducing lymphocyte sequestration and immunosuppression^{17,27}. Therapeutic augmentation of CD69 levels may be a new method for regulating S1P₁ function in lymphoid cells.

METHODS

Mice, adoptive transfers and treatments. C57BL/6 (B6) and B6-CD45.1 mice were obtained from the National Cancer Institute. *Ifnar1*^{-/-} mice²⁸ were ten generations and CD69^{-/-} mice¹¹ six generations backcrossed to B6 mice. S1P₁^{-/-} thymocytes and B cells were generated as described previously⁶. For adoptive transfers, spleen and lymph node cells were labelled with either 1 μM 5-(and-6)-(((4-chloromethyl)benzoyl)amino)tetramethylrhodamine (CMTMR; Molecular Probes) or 1 μM 5-(and-6)-carboxyfluorescein diacetate succinimidyl ester

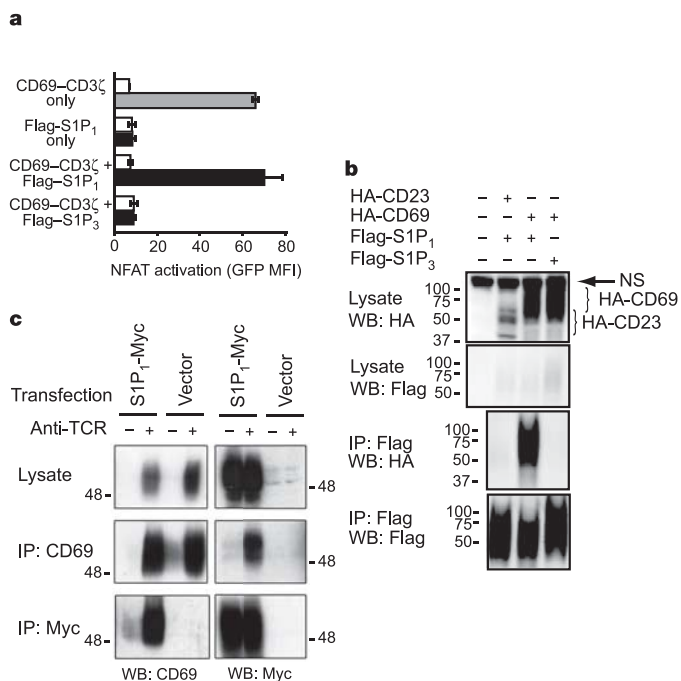


Figure 4 | Mouse and human CD69 interact with S1P₁. **a**, 2B4 NFAT-GFP reporter cells were singly transduced or co-transduced as indicated. Receptors were crosslinked with plate-bound antibodies (open columns, isotype control; grey column, anti-CD69; black columns, anti-Flag) and the expression of GFP, shown as geometric mean fluorescence intensity (MFI), was analysed 16 h later. Data are shown as means ± s.d. and are pooled from three experiments. **b**, Lysates of WEHI-231 cells transduced with the indicated retroviral constructs were immunoprecipitated (IP) with anti-Flag. Coimmunoprecipitated proteins were detected by using anti-haemagglutinin (anti-HA) or anti-Flag antibodies. Data shown are representative of four independent experiments. NS, non-specific band; WB, western blot. **c**, Lysates of Jurkat cells transfected with S1P₁-Myc or an empty vector and stimulated or not with anti-TCR antibody C305 to upregulate CD69 expression were immunoprecipitated with anti-CD69 or anti-Myc antibodies. Coimmunoprecipitated proteins were detected on western blots by using anti-CD69 and anti-Myc antibodies. Data shown are representative of three independent experiments. All markers are in kDa.

(CFSE; Molecular Probes)²⁹ and mixed together or with wild-type control CD45.1 lymphocytes and injected intravenously into recipients. After 36–40 h, mice were injected intravenously with 100 μg of poly(I:C) (Sigma) or an equivalent volume of saline. Treatment with poly(I:C) was for 6 h unless otherwise indicated. For infection with LCMV, 2 × 10⁶ plaque-forming units of Armstrong strain were injected intravenously 24 h after adoptive transfer. Lymph collection was as described previously⁶. Statistical analysis was with a paired two-tailed Student's *t*-test.

Flow cytometry, transwell migration assays, cell purification and quantitative PCR. Procedures were similar to those described previously^{6,15,29} and are detailed in Supplementary Methods.

Constructs and retroviral transduction. Details on MSCV2.2 retroviral constructs and transduction procedures are provided in Supplementary Methods. Quantitative PCR analysis of total S1P₁ and CD69 transcript abundance in two preparations of co-transduced WEHI231 cells showed they were within fivefold of the levels in T cells isolated from mice 6 h after treatment with poly(I:C).

Reporter cell assay. 2B4 cells containing an NFAT-GFP reporter were transduced and analysed as described previously^{20,30}.

Immunoprecipitation. Transduced WEHI231 cells were lysed in 0.875% Brij 97, 0.125% Nonidet P40, 150 mM NaCl, 10 mM Tris-HCl pH 7.4, 0.02% NaN₃ buffer containing protease inhibitors (Sigma). Transfected Jurkat cells were stimulated or not with anti-Jurkat TCR Vβ antibody C305 (ascites diluted 1:1,000) and lysed with 1% n-dodecyl-β-D-maltoside (Calbiochem) in 20 mM Tris-HCl pH 7.5 containing 100 mM NaCl, 10 mM EDTA, 50 mM NaF and 1 mM Pefabloc. Immunoprecipitation was by standard procedures as described in Supplementary Methods.

Received 6 December 2005; accepted 26 January 2006.

Published online 8 March 2006.

- Hall, J. G. & Morris, B. The immediate effect of antigens on the cell output of a lymph node. *Br. J. Exp. Pathol.* **46**, 450–454 (1965).
- Degre, M. Influence of polyinosinic:polycytidylic acid on the circulating white blood cells in mice. *Proc. Soc. Exp. Biol. Med.* **142**, 1087–1091 (1973).
- Korngold, R., Blank, K. J. & Murasko, D. M. Effect of interferon on thoracic duct lymphocyte output: induction with either poly I:poly C or vaccinia virus. *J. Immunol.* **130**, 2236–2240 (1983).
- Kalaaji, A. N., Abernethy, N. J., McCullough, K. & Hay, J. B. Recombinant bovine interferon- α 1 inhibits the migration of lymphocytes from lymph nodes but not into lymph nodes. *Reg. Immunol.* **1**, 56–61 (1988).
- Young, A. J., Seabrook, T. J., Marston, W. L., Dudler, L. & Hay, J. B. A role for lymphatic endothelium in the sequestration of recirculating gamma delta T cells in TNF- α -stimulated lymph nodes. *Eur. J. Immunol.* **30**, 327–334 (2000).
- Matloubian, M. *et al.* Lymphocyte egress from thymus and peripheral lymphoid organs is dependent on S1P receptor 1. *Nature* **427**, 355–360 (2004).
- Allende, M. L., Dreier, J. L., Mandala, S. & Proia, R. L. Expression of the sphingosine-1-phosphate receptor, S1P1, on T-cells controls thymic emigration. *J. Biol. Chem.* **279**, 15396–15401 (2004).
- Feng, C. *et al.* A potential role for CD69 in thymocyte emigration. *Int. Immunol.* **14**, 535–544 (2002).
- Nakayama, T. *et al.* The generation of mature, single-positive thymocytes *in vivo* is dysregulated by CD69 blockade or overexpression. *J. Immunol.* **168**, 87–94 (2002).
- Lauzurica, P. *et al.* Phenotypic and functional characteristics of hematopoietic cell lineages in CD69-deficient mice. *Blood* **95**, 2312–2320 (2000).
- Murata, K. *et al.* CD69-null mice protected from arthritis induced with anti-type II collagen antibodies. *Int. Immunol.* **15**, 987–992 (2003).
- Testi, R., D'Ambrosio, D., De Maria, R. & Santoni, A. The CD69 receptor: a multipurpose cell-surface trigger for hematopoietic cells. *Immunol. Today* **15**, 479–483 (1994).
- Sancho, D., Gomez, M. & Sanchez-Madrid, F. CD69 is an immunoregulatory molecule induced following activation. *Trends Immunol.* **26**, 136–140 (2005).
- Risso, A. *et al.* CD69 in resting and activated T lymphocytes. Its association with a GTP binding protein and biochemical requirements for its expression. *J. Immunol.* **146**, 4105–4114 (1991).
- Cinamon, G. *et al.* Sphingosine 1-phosphate receptor 1 promotes B cell localization in the splenic marginal zone. *Nature Immunol.* **5**, 713–720 (2004).
- Chu, P. *et al.* Systematic identification of regulatory proteins critical for T-cell activation. *J. Biol.* **2**, 21.1–21.16 (2003).
- Brinkmann, V., Cyster, J. G. & Hla, T. FTY720: sphingosine 1-phosphate receptor-1 in the control of lymphocyte egress and endothelial barrier function. *Am. J. Transplant.* **4**, 1019–1025 (2004).
- Rosen, H., Alfonso, C., Surh, C. D. & McHeyzer-Williams, M. G. Rapid induction of medullary thymocyte phenotypic maturation and egress inhibition by nanomolar sphingosine 1-phosphate receptor agonist. *Proc. Natl Acad. Sci. USA* **100**, 10907–10912 (2003).
- Schwab, S. R. *et al.* Lymphocyte sequestration through S1P lyase inhibition and disruption of S1P gradients. *Science* **309**, 1735–1739 (2005).
- Arase, H., Mocarski, E. S., Campbell, A. E., Hill, A. B. & Lanier, L. L. Direct recognition of cytomegalovirus by activating and inhibitory NK cell receptors. *Science* **296**, 1323–1326 (2002).
- Lopez-Cabrera, M. *et al.* Transcriptional regulation of the gene encoding the human C-type lectin leukocyte receptor AIM/CD69 and functional characterization of its tumor necrosis factor- α -responsive elements. *J. Biol. Chem.* **270**, 21545–21551 (1995).
- Graeler, M. & Goetzl, E. J. Activation-regulated expression and chemotactic function of sphingosine 1-phosphate receptors in mouse splenic T cells. *FASEB J.* **16**, 1874–1878 (2002).
- Esplugues, E. *et al.* Enhanced antitumor immunity in mice deficient in CD69. *J. Exp. Med.* **197**, 1093–1106 (2003).
- Ramirez, R., Carracedo, J., Zamzami, N., Castedo, M. & Kroemer, G. Pertussis toxin inhibits activation-induced cell death of human thymocytes, pre-B leukemia cells and monocytes. *J. Exp. Med.* **180**, 1147–1152 (1994).
- Sancho, D. *et al.* Functional analysis of ligand-binding and signal transduction domains of CD69 and CD23 C-type lectin leukocyte receptors. *J. Immunol.* **165**, 3868–3875 (2000).
- Bikah, G., Pogue-Caley, R. R., McHeyzer-Williams, L. J. & McHeyzer-Williams, M. G. Regulating T helper cell immunity through antigen responsiveness and calcium entry. *Nature Immunol.* **1**, 402–412 (2000).
- Rosen, H. & Liao, J. Sphingosine 1-phosphate pathway therapeutics: a lipid ligand-receptor paradigm. *Curr. Opin. Chem. Biol.* **7**, 461–468 (2003).
- Muller, U. *et al.* Functional role of type I and type II interferons in antiviral defense. *Science* **264**, 1918–1921 (1994).
- Lo, C. G., Xu, Y., Proia, R. L. & Cyster, J. G. Cyclical modulation of sphingosine-1-phosphate receptor 1 surface expression during lymphocyte recirculation and relationship to lymphoid organ transit. *J. Exp. Med.* **201**, 291–301 (2005).
- Rosen, D. B. *et al.* Cutting edge: Lectin-like transcript-1 is a ligand for the inhibitory human NKR-PIA receptor. *J. Immunol.* **175**, 7796–7799 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Nakayama and S. Zeigler for CD69^{-/-} mice; K. Murali-Krishna for *Ifnar1*^{-/-} mice; R. Proia for *S1p1*^{-/-} mice; S. Jahn for technical assistance; C. Allen, S. Schwab and C. McArthur for help with cell sorting; K. Kabashima for fetal liver chimeras; O. Lam for mouse husbandry; and S. Schwab and C. Lo for comments on the manuscript. L.R.S. is supported by the UCSF Medical Scientist Training Program; D.B.R. is supported by the Genentech Graduate Fellowship; M.M. was supported by the Pfizer Postdoctoral Fellowship in Immunology and Rheumatology, a research award from the Arthritis Foundation, the Rosalind Russell Medical Research Center for Arthritis at UCSF, and the Sandler Family Supporting Foundation; L.L.L. is an American Cancer Society Research Professor; J.G.C. is an Investigator of the Howard Hughes Medical Institute. This work was supported in part by grants from the NIH (to L.L.L. and J.G.C.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.G.C. (jason.cyster@ucsf.edu) or M.M. (mehrdad.matloubian@ucsf.edu).

An excitable gene regulatory circuit induces transient cellular differentiation

Gürol M. Süel¹, Jordi Garcia-Ojalvo², Louisa M. Liberman¹ & Michael B. Elowitz¹

Certain types of cellular differentiation are probabilistic and transient^{1–3}. In such systems individual cells can switch to an alternative state and, after some time, switch back again. In *Bacillus subtilis*, competence is an example of such a transiently differentiated state associated with the capability for DNA uptake from the environment. Individual genes and proteins underlying differentiation into the competent state have been identified^{4,5}, but it has been unclear how these genes interact dynamically in individual cells to control both spontaneous entry into competence and return to vegetative growth. Here we show that this behaviour can be understood in terms of excitability in the underlying genetic circuit. Using quantitative fluorescence time-lapse microscopy, we directly observed the activities of multiple circuit components simultaneously in individual cells, and analysed the resulting data in terms of a mathematical model. We find that an excitable core module containing positive and negative feedback loops can explain both entry into, and exit from, the competent state. We further tested this model by analysing initiation in sister cells, and by re-engineering the gene circuit to specifically block exit. Excitable dynamics driven by noise naturally generate stochastic and transient responses⁶, thereby providing an ideal mechanism for competence regulation.

Upon encountering nutrient limitation, a minority of *B. subtilis* cells become competent for DNA uptake while most commit irreversibly to sporulation (Fig. 1a). Extensive research has elucidated a detailed map of molecular interactions that comprise the *B. subtilis* competence control circuit (Supplementary Fig. S1)^{4,5}. At the heart of this circuit is the ComK ‘master’ transcription factor (Fig. 1b). ComK activates expression of a suite of genes necessary for competence, including the *comG* operon (Fig. 1b)^{7–10}. ComK also activates its own expression, and two recent studies have shown that this positive feedback loop is critical for the induction of competence^{11,12}. Upon entry into stationary phase, *comK* is expressed at a basal level, but is also rapidly degraded by the MecA complex, a multiprotein assembly that includes the ClpP–ClpC proteases¹³. Independent studies have shown that the ComS peptide competitively inhibits ComK degradation by the MecA complex^{13,14}. Expression of ComS thus favours induction of competence by allowing ComK levels to build up sufficiently to enable full ComK activation by positive autoregulation. Despite the important role of ComS in inducing competence, its expression is activated by stress and cell–cell signalling and is therefore high in all cells under stress, not just those destined to become competent⁷. Interestingly, overexpression of ComK was suggested to suppress expression of *comS*⁷. This implies the existence of an indirect negative feedback loop acting upon ComK, which might affect exit from competence (Fig. 1b). Therefore, we consider here interactions among the MecA complex, *comK* and *comS*, which we collectively refer to as the ‘MeKS’ module.

Differentiation into the competent state is difficult to study by

traditional methods that average over large populations: the process affects a small minority of cells in a non-synchronous manner. Here we built strains in which activities of different pairs of promoters within the circuit could be monitored simultaneously in the same cell, revealing interactions between circuit components that could not be obtained by observing the same genes one at a time. Pairwise combinations of the promoters P_{comG} , P_{comK} and P_{comS} expressing *yfp* or *cfp* were inserted into standard sites within the *B. subtilis* chromosome, leaving the endogenous genes intact (see Supplementary Information). We analysed these strains using automated time-lapse fluorescence microscopy, and quantitative image analysis¹⁵, under conditions where the overall frequency of competence events was $3.6 \pm 0.7\%$ (26 out of 725 cell division events generated competent cells).

The resulting data can be interpreted in terms of a mathematical model of the MeKS module. As described in Box 1, the MeKS model is a dynamical system that exhibits excitability: relatively small, threshold-crossing perturbations trigger large-amplitude excursions in phase space that eventually return the system to its initial state⁶. Correspondingly, in the cell, stochastic effects in gene expression have been shown to generate significant variability^{16,17}. Such biochemical ‘noise’ can initiate a sequence of intracellular events—changes in gene expression and protein degradation—that cause cells to transiently differentiate into the competent state and subsequently return to vegetative growth. Noise-driven excitable systems naturally exhibit the two key characteristics of competence: probabilistic and transient activation.

To explore these properties further, we first simultaneously measured P_{comG} and P_{comK} promoter activities in single competent cells. P_{comG} and P_{comK} are both regulated by ComK, but P_{comK} also has many other important transcriptional inputs (Supplementary Fig. S1). This experiment evaluates the significance of these other transcriptional inputs into P_{comK} during competence. Figure 2a shows frames from recorded film footage of a *B. subtilis* microcolony under nutrient limitation conditions, in which most cells eventually sporulate (see also Supplementary Movie 1). In this example, a single cell within the microcolony becomes competent and consequently its cell division is blocked, as previously reported¹⁸. As P_{comG} and P_{comK} activities decline during exit from competence, the elongated cell undergoes multiple septation events and returns to the vegetative state. Quantitative time traces of the fluorescence levels show strong correlation between P_{comG} and P_{comK} in this film footage (Fig. 2b). This behaviour is representative of all measured events in this strain ($n = 37$) (Fig. 2c). The high degree of correlation between P_{comG} and P_{comK} suggests that other transcriptional inputs to the ComK promoter do not significantly affect ComK expression during competence.

In the MeKS model, ComK indirectly represses ComS, generating an anti-correlation between P_{comG} and P_{comS} activities. The regulation

¹Division of Biology and Department of Applied Physics, California Institute of Technology, Pasadena, California 91125, USA. ²Departament de Física i Enginyeria Nuclear, Universitat Politècnica de Catalunya, Colom 11, E-08222 Terrassa, Spain.

of *comS* is, however, known to be complex, having several transcriptional inputs (Supplementary Fig. S1)^{19–21}. To test the prediction of the MeKS model we constructed a strain containing copies of the P_{comG} and P_{comS} promoters expressing *cfp* and *yfp*, respectively. As shown in Fig. 3a and Supplementary Movie 2, all cells express P_{comS} to varying degrees. In cells that become competent, P_{comG} activity increases as P_{comS} activity decreases. Later, as P_{comG} activity shuts off and septation begins, P_{comS} activity increases again. This striking negative correlation between P_{comG} and P_{comS} activities is present during both entry and exit from competence, although it is more closely synchronized during entry (Fig. 3b). This behaviour is representative of data obtained from all competent cells of the same strain ($n = 31$) (Fig. 3c), and is consistent with negative

regulation of *comS* by ComK. Furthermore, the negative correlation is specific to competence, as is evident from the behaviour of the non-competent sister cell in Fig. 3b.

A fundamental question is whether initiation of competence is stochastic or affected by memory of previous events. Escape from competence returns promoter activities to pre-competence levels, suggesting the possibility of successive episodes of competence. Indeed, as shown in Fig. 3d, two consecutive competence events can be observed in a single cell lineage, showing that cells retain the potential to re-initiate competence. In fact, re-initiation occurred with a frequency of $6.0 \pm 2.0\%$ ($n = 9$ events out of 151), not significantly different from the overall competence frequency ($3.6 \pm 0.7\%$). Repeated competence events are neither favoured nor suppressed. This evidence for stochastic initiation of competence is further supported by analysis of competence events in sister cell pairs (Supplementary Information). Cells were not significantly more or less likely to become competent if their sister became competent (conditional frequency = $4.1 \pm 0.9\%$, $n = 19$ events out of 463). When two sisters do become competent together, the amount of time one spends in competence is uncorrelated with that of its sister cell (Kolmogorov–Smirnov test; $n = 36$). These results are consistent with a stochastic and memory-less model for competence initiation and duration.

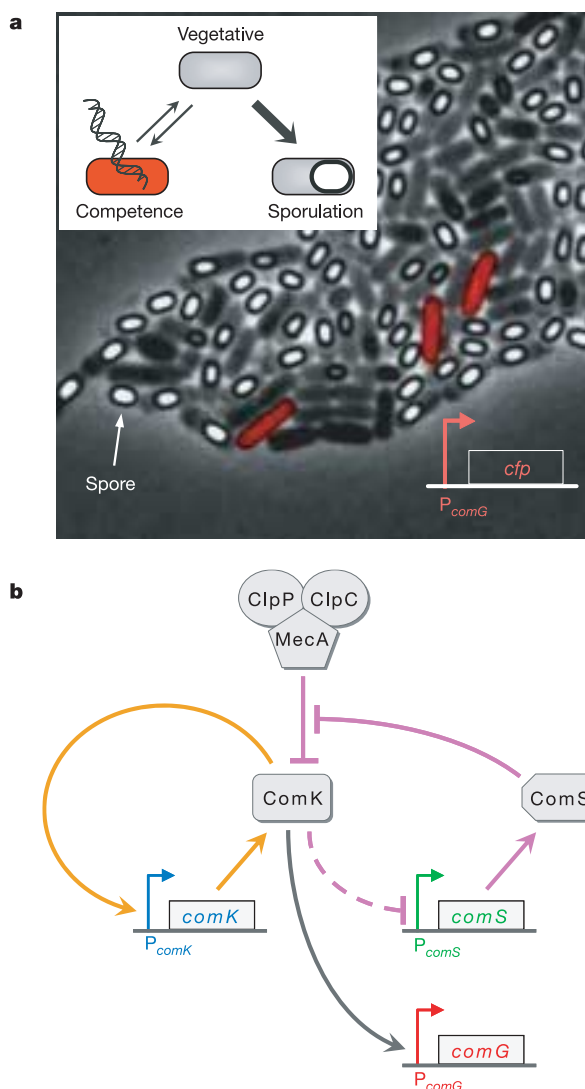


Figure 1 | Stress response in *B. subtilis* and the core competence circuit.

a, Snapshot of a *B. subtilis* microcolony in nutrient-limited conditions. *cfp* expression from P_{comG} is shown in red. Inset: a flow chart illustrating developmental paths connecting the vegetative, spore forming and competent states. **b**, Map of interactions within the core competence circuit (MeKS). The transcriptional autoregulatory positive feedback loop of ComK and the ComS-mediated indirect negative feedback loop are depicted in orange and purple, respectively. ComS competes with ComK for degradation by the MecA–ClpP–ClpC complex, effectively interfering with degradation of ComK (curved purple inhibitory arrow). The dashed purple line from ComK to P_{comS} denotes indirect repression. The activities of the promoters labelled in red, blue and green were measured in this study. These colours are used to represent the corresponding promoters throughout the figures.

Box 1 | The dynamical model of competence induction

To understand how the MeKS network structure determines the dynamics of competence, we built a mathematical model constrained by experimental observations (see Supplementary Information). This model can be reduced to a system of two stochastic ordinary differential equations incorporating both the direct positive and the ComS-mediated negative feedback loops of ComK. In dimensionless form:

$$\frac{dK}{dt} = a_k + \frac{b_k K^n}{K_0^n + K^n} - \frac{K}{1 + K + S} \quad (1)$$

$$\frac{dS}{dt} = \frac{b_s}{1 + (K/k_1)^p} - \frac{S}{1 + K + S} + \xi(t) \quad (2)$$

Here, K and S represent the concentration levels of ComK and ComS protein, respectively. a_k and b_k represent minimal and fully activated rates of ComK production, respectively. K_0 is the concentration of ComK required for 50% activation. The cooperativities of ComK auto-activation and ComS repression are parameterized by the Hill coefficients n and p , respectively. Expression of ComS has maximum rate b_s and is half-maximal when $K = k_1$. Enzymatic MecA-mediated degradation affects both ComK and ComS; the form of the corresponding nonlinear degradation terms expresses a competitive mechanism, which is the only source of coupling from ComS to ComK. Random fluctuations in ComS expression are represented by a noise term $\xi(t)$ (see Supplementary Information for a more detailed analysis).

The dynamical behaviour of equations (1) and (2) without noise can be analysed graphically by plotting their nullclines and vector field in the ComK–ComS phase space (Fig. 4a) for appropriate parameters (given in the Supplementary Information). This analysis reveals three fixed points: a stable node at low ComK (the vegetative state) and two unstable fixed points. Of these, the one at intermediate ComK is an unstable saddle and the one at high ComK (the competent state) is an unstable spiral. No limit-cycle behaviour coexists with the stable vegetative state in this parameter region (see Supplementary Information). Under these conditions, the system is capable of excitable behaviour: relatively small perturbations from the vegetative state may cause long excursions through phase space around the unstable spiral at high ComK (that is, through the competence region), as determined by the vector field. The vegetative state can be perturbed by noise in the expression of either ComK or ComS, leading to these transient differentiation events. Samples of such trajectories, generated by numerical integration of the model, are superimposed as pink lines in Fig. 4a and plotted against time in Fig. 4b.

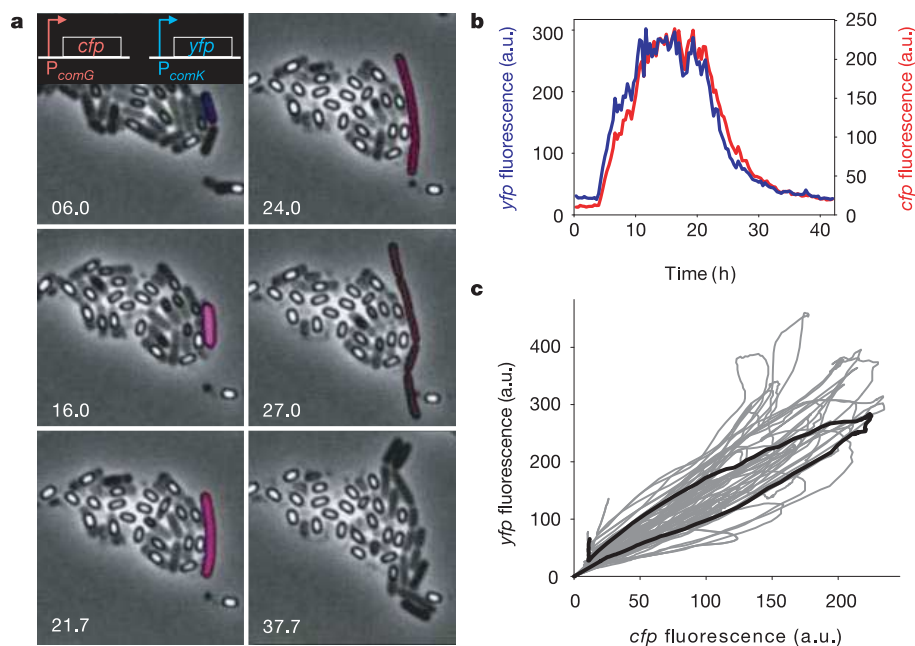


Figure 2 | Activities of P_{comK} and P_{comG} promoters are highly correlated during competence. **a**, Frames from film footage of a typical competence event. *yfp* expressed from P_{comK} and *cfp* expressed from P_{comG} are coloured blue and red, respectively (see also Supplementary Movie 1). Overlapping dynamics of P_{comK} and P_{comG} result in the purple (red plus blue) colour of the competent cell. Time (in hours) is indicated for each frame. **b**, Quantitative time series of P_{comK} -*yfp* (blue line) and P_{comG} -*cfp* (red line)

obtained through semi-automated data processing of the competence event shown in **a**. P_{comK} and P_{comG} activities exhibit nearly identical dynamics. a.u., arbitrary units. **c**, Plot of P_{comK} -*yfp* versus P_{comG} -*cfp* obtained from all ($n = 37$) competence events in this strain. The traces are smoothed for visual clarification. Note the positive correlation between P_{comK} -*yfp* and P_{comG} -*cfp*. Highlighted in black is the competence time trace depicted in **b**.

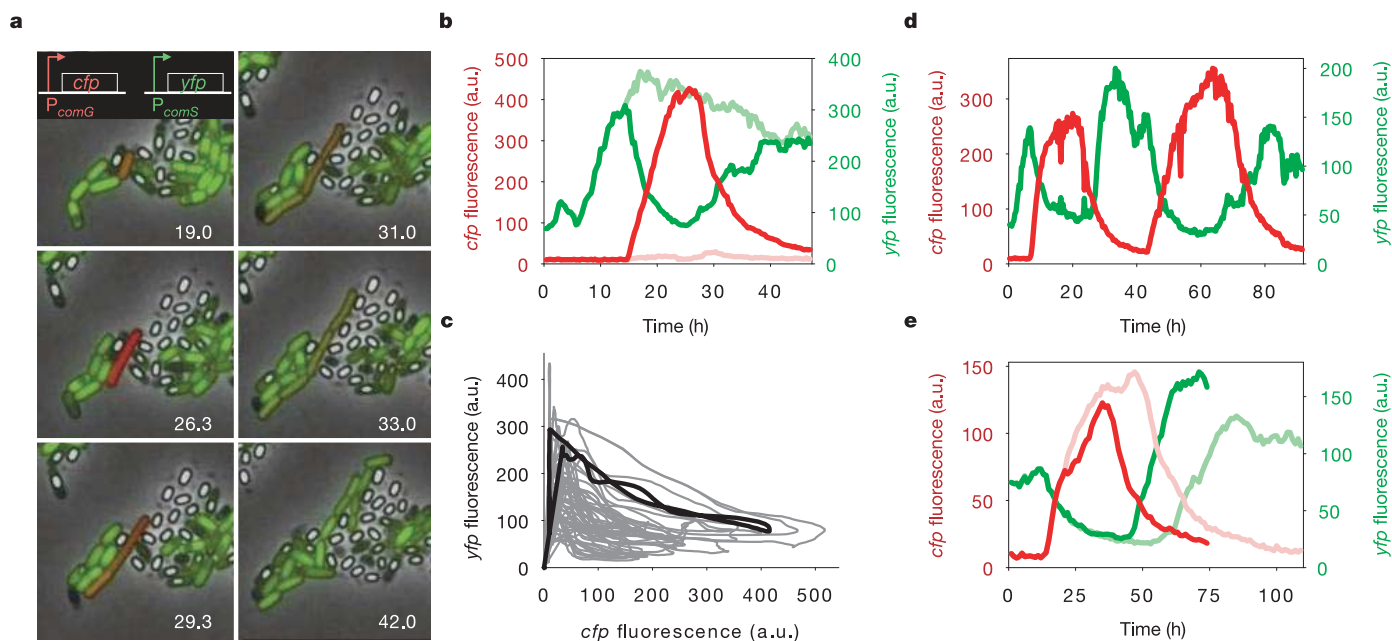


Figure 3 | Promoter activities of P_{comS} and P_{comG} are anti-correlated during competence. **a**, Frames from film footage of a typical competence event with P_{comS} -*yfp* and P_{comG} -*cfp* expression shown in green and red, respectively (see inset) (see also Supplementary Movie 2). Time (in hours) is indicated for each frame. **b**, Quantitative time series of P_{comS} -*yfp* (green line) and P_{comG} -*cfp* (red lines) for the competence event shown in **a**. Depicted in faint green and faint red are P_{comS} and P_{comG} activities obtained from the non-competent sister cell. Note that the negative correlation between P_{comS} and P_{comG} expression dynamics is only observed in the competent cell.

c, Plot of P_{comS} -*yfp* versus P_{comG} -*cfp* obtained from all ($n = 31$) competence events in this strain. The traces are smoothed for visual clarification. Note the negative correlation between P_{comS} -*yfp* and P_{comG} -*cfp*. Highlighted in black is the competence time trace depicted in **b**. **d**, Quantitative time traces of consecutive competence events within a single cell lineage. The colour scheme is identical to that described for **b**. **e**, Quantitative time traces of P_{comS} -*yfp* and P_{comG} -*cfp* measured in sister cells that both undergo competence. The colour scheme is identical to that described for **b**. Note the different durations of competence in sister cells (see Supplementary Fig. S7).

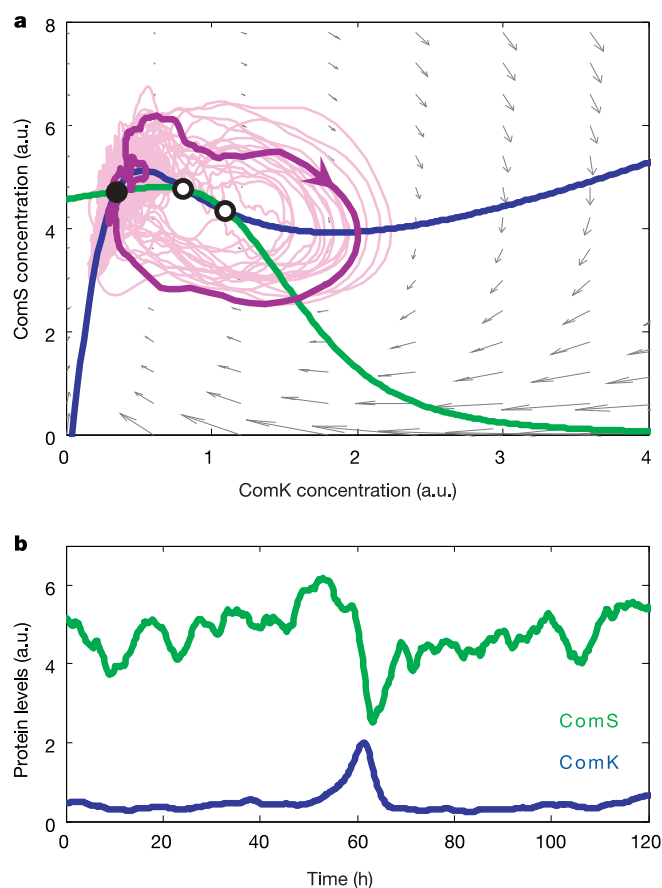


Figure 4 | Modelling of the core competence network reveals an excitable system. **a**, Phase plane diagram formed by the system of equations shown in Box 1. Nullclines for equations (1) and (2) are shown in blue and green, respectively. Grey arrows represent the vector field of the dynamical system. The stable steady-state corresponding to vegetative growth is indicated with a black filled circle. The saddle and the unstable competent fixed points are indicated with open circles. A set of excursion trajectories is shown in pink, with a single representative trajectory of the system highlighted in purple. Initiation of excursions in phase space is triggered by noise (Box 1), and trajectories are determined by the phase space vector field. **b**, Simulations of ComS (green) and ComK (blue) activities as a function of time. Note the negative correlation between the ComS and ComK levels during competence, consistent with experimental observations.

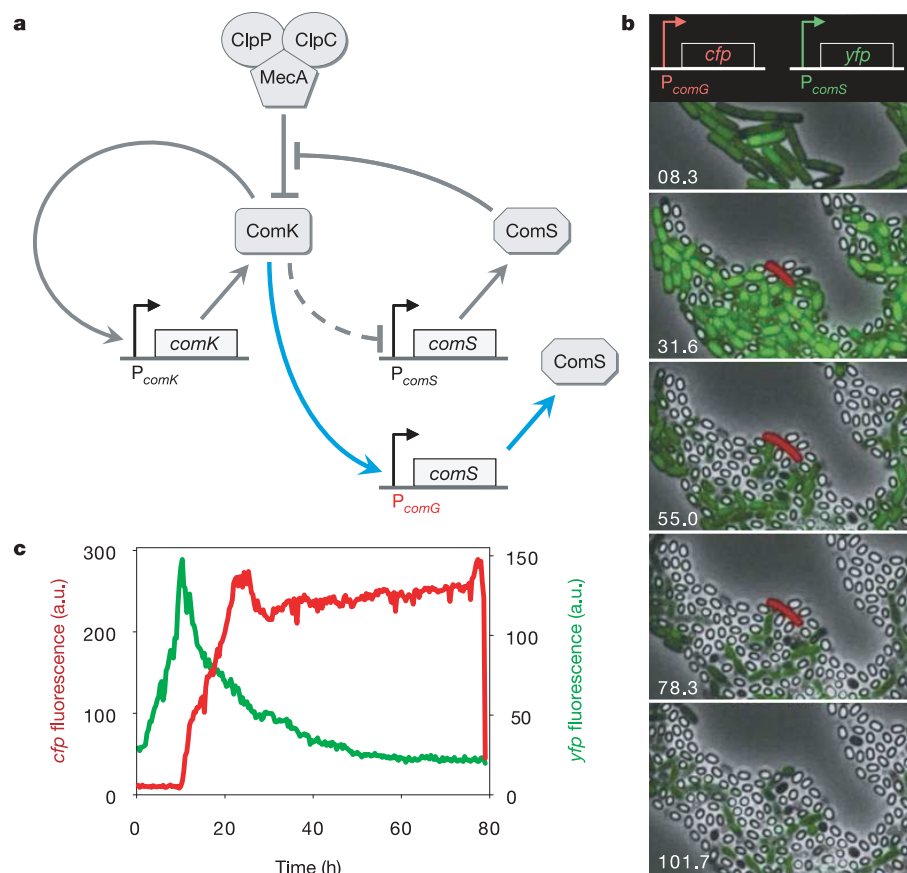


Figure 5 | Competence lock through feedback bypass. **a**, Network schematic depicting in blue the extra link introduced to bypass the native ComS-mediated negative feedback loop (FeBy strain). Note that the native network is left intact (see Fig. 1a). **b**, Frames from film footage of a typical competence event in the FeBy strain, with P_{comS} and P_{comG} activities depicted in green and red, respectively (see also Supplementary Movie 3). **c**, Quantitative time series of P_{comS} -yfp (green line) and P_{comG} -cfp (red line) for the event shown in **b**. The FeBy cell enters competence, but cannot exit from competence and eventually lyses (note the sudden drop after nearly 80 h).

To understand how the network structure supports the dynamics of competence, we built a differential equation model of the MeKS module that incorporates noise. In this model, two feedback loops regulate ComK: the positive loop is based on direct transcriptional auto-activation; the negative loop is indirect and involves ComS-modulated degradation of ComK (Fig. 1a, Box 1 and Supplementary Information). This model contains a regime in which noise in the production of either ComK or ComS can trigger large excursions in phase space that resemble experimental observations (Fig. 4). Thus, this model of excitability accounts in a natural way for the dynamics of competence observed in experiments.

How does the circuit generate transient competence events? This behaviour is caused by the combination of a fast positive and slow negative feedback. Positive feedback motifs are well known generators of bistability in cell biology^{22,23}. Combinations of positive and negative feedbacks are crucial elements of dynamic processes in biology such as the cell cycle²⁴. In the present system, noise can induce escape from the otherwise stable vegetative state and turn the system 'on' via the ComK positive feedback, as suggested previously by other groups^{11,12}. On a slower timescale, this initiates the ComS-mediated negative feedback. Reduction in ComS levels eventually shuts the system back 'off' through an increase in ComK degradation, returning the cell to its vegetative state. In this way, ComS has a dual role in the system: on the one hand it is necessary to initiate competence, by blocking degradation of ComK and allowing positive autoregulation to take effect; on the other hand, repression of ComS is necessary for exit from competence, because reduction in ComS levels favours ComK degradation by MecA.

The excitable model of competence induction makes a strong prediction: interfering with the ComS-mediated negative feedback loop should stabilize the competent state without changing the frequency with which cells become competent (Supplementary Fig. S6). To test this prediction, we constructed a feedback bypass (FeBy) strain of *B. subtilis* containing a single chromosomal copy of *P_{comG}* driving the expression of *comS* (Fig. 5a). This modification leaves the native competence network intact: the added *comS* gene is expressed only in competent cells. By expressing additional ComS in this way, the natural negative feedback is rendered ineffective.

We performed time-lapse microscopy on the FeBy strain (Fig. 5b and Supplementary Movie 3). The resulting movies show that FeBy cells become competent at normal frequencies ($4.2 \pm 0.8\%$; $n = 32$ out of 754), but most of these competent cells fail to return to vegetative growth. In a wild-type culture, 61% of cells successfully exit competence ($n = 136$), compared to only 12% of FeBy cells ($n = 55$). *P_{comG}* expression levels in competent FeBy cells do not significantly exceed those in wild-type cells, suggesting that the defect in exit is not simply due to overexpression of ComK. As in wild-type cells, a negative correlation between *P_{comG}* and *P_{comS}* is observed in FeBy cells (Fig. 5c). Thus, the FeBy cells behave similarly to wild-type cells during entry into competence, and differ specifically in their reduced ability to exit.

Even in well-characterized cellular networks, it is generally difficult to connect circuit structure to circuit dynamics²⁵. Within the complex *B. subtilis* competence circuitry however, the relatively simple MeKS module can explain the transient and probabilistic features of competence dynamics. This suggests that during competence not all of the known interactions within the larger competence circuit are important all of the time.

In the FeBy strain, a single rewiring step transformed a transient differentiation process into a terminal one. This reflects a degree of plasticity in the excitable circuit module. Notably, the architecture of this circuit is similar to the circadian clock and cell cycle oscillators, which also contain combined positive–negative feedback loops^{24,26,27}. Excitability is found in biological systems that process information such as neurons, where the frequency of events, but not their shape or amplitude, needs to be tuned²⁸. Our results show that excitability also operates in an intracellular genetic network controlling transient

differentiation—an event on a very different timescale that uses completely different molecular components. Evidently, analogous functional roles for excitable dynamics exist in microbial and neuronal systems. We anticipate that this circuit design may therefore represent a general solution to a range of biological design problems.

METHODS

For detailed information on all methods, see Supplementary Information

Strain construction. Promoter–*yfp/cfp* fusions were generated using fusion polymerase chain reaction techniques and cloned into *B. subtilis* integration vectors pDL30 (gift from J. Dworkin), pPyr-Cm and pSac-Cm (constructed by R. Middleton). The strain background was *B. subtilis* strain PY79.

Preparation of cells for microscopy. Cells were grown at 37 °C in Luria broth to an optical density of 1.8 and re-suspended in 0.5 volume of resuspension medium (RM) supplemented with 0.02% glucose. After 1.5 h incubation at 37 °C, cells were applied onto a 1.5% agarose pad made with RM and placed into a coverslip-bottom Willco dish for imaging.

Time-lapse microscopy. Growth of microcolonies was observed with fluorescence time-lapse microscopy at 37 °C using an Olympus IX-81 inverted microscope with a motorized stage (ASI). Image sets were acquired every 20 min with a Hamamatsu ORCA-ER camera. Custom Visual Basic software was used to automate image acquisition and microscope control.

Data analysis. Time-lapse microscopy data were analysed with custom software written in Matlab.

Model analysis. Mathematical models were simulated using custom-made software and further analysed using the analytic continuation software packages AUTO and DDE-BIFTOOL.

Received 15 September 2005; accepted 18 January 2006.

- Dupin, E., Real, C., Glavieux-Pardanaud, C., Vaigot, P. & Le Douarin, N. M. Reversal of developmental restrictions in neural crest lineages: transition from Schwann cells to glial-melanocytic precursors *in vitro*. *Proc. Natl Acad. Sci. USA* **100**, 5229–5233 (2003).
- Lee, E. J., Russell, T., Hurley, L. & Jameson, J. L. Pituitary transcription factor-1 induces transient differentiation of adult hepatic stem cells into prolactin-producing cells *in vivo*. *Mol. Endocrinol.* **19**, 964–971 (2005).
- Meeks, J. C., Campbell, E. L., Summers, M. L. & Wong, F. C. Cellular differentiation in the cyanobacterium *Nostoc punctiforme*. *Arch. Microbiol.* **178**, 395–403 (2002).
- Grossman, A. D. Genetic networks controlling the initiation of sporulation and the development of genetic competence in *Bacillus subtilis*. *Annu. Rev. Genet.* **29**, 477–508 (1995).
- Hamoen, L. W., Venema, G. & Kuipers, O. P. Controlling competence in *Bacillus subtilis*: shared use of regulators. *Microbiology* **149**, 9–17 (2003).
- Lindner, B., Garcia-Ojalvo, J., Neiman, A. & Schimansky-Geier, L. Effects of noise in excitable systems. *Phys. Rep.* **392**, 321–424 (2004).
- Hahn, J., Kong, L. & Dubnau, D. The regulation of competence transcription factor synthesis constitutes a critical control point in the regulation of competence in *Bacillus subtilis*. *J. Bacteriol.* **176**, 5753–5761 (1994).
- Hahn, J., Luttinger, A. & Dubnau, D. Regulatory inputs for the synthesis of ComK, the competence transcription factor of *Bacillus subtilis*. *Mol. Microbiol.* **21**, 763–775 (1996).
- van Sinderen, D. *et al.* *comK* encodes the competence transcription factor, the key regulatory protein for competence development in *Bacillus subtilis*. *Mol. Microbiol.* **15**, 455–462 (1995).
- van Sinderen, D. & Venema, G. *comK* acts as an autoregulatory control switch in the signal transduction route to competence in *Bacillus subtilis*. *J. Bacteriol.* **176**, 5762–5770 (1994).
- Maamar, H. & Dubnau, D. Bistability in the *Bacillus subtilis* K-state (competence) system requires a positive feedback loop. *Mol. Microbiol.* **56**, 615–624 (2005).
- Smits, W. K. *et al.* Stripping *Bacillus*: ComK auto-stimulation is responsible for the bistable response in competence development. *Mol. Microbiol.* **56**, 604–614 (2005).
- Turgay, K., Hahn, J., Burghoorn, J. & Dubnau, D. Competence in *Bacillus subtilis* is controlled by regulated proteolysis of a transcription factor. *EMBO J.* **17**, 6730–6738 (1998).
- Ogura, M., Liu, L., Lacelle, M., Nakano, M. M. & Zuber, P. Mutational analysis of ComS: evidence for the interaction of ComS and MecA in the regulation of competence development in *Bacillus subtilis*. *Mol. Microbiol.* **32**, 799–812 (1999).
- Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307**, 1962–1965 (2005).
- Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genet.* **31**, 69–73 (2002).
- Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).

18. Hajjema, B. J., Hahn, J., Haynes, J. & Dubnau, D. A. ComGA-dependent checkpoint limits growth during the escape from competence. *Mol. Microbiol.* **40**, 52–64 (2001).
19. Serror, P. & Sonenshein, A. L. CodY is required for nutritional repression of *Bacillus subtilis* genetic competence. *J. Bacteriol.* **178**, 5910–5915 (1996).
20. Nakano, M. M. & Zuber, P. The primary role of *comA* in establishment of the competent state in *Bacillus subtilis* is to activate expression of *srfA*. *J. Bacteriol.* **173**, 7269–7274 (1991).
21. Hahn, J. & Dubnau, D. Growth stage signal transduction and the requirements for *srfA* induction in development of competence. *J. Bacteriol.* **173**, 7275–7282 (1991).
22. Isaacs, F. J., Hasty, J., Cantor, C. R. & Collins, J. J. Prediction and measurement of an autoregulatory genetic module. *Proc. Natl Acad. Sci. USA* **100**, 7714–7719 (2003).
23. Xiong, W. & Ferrell, J. E. Jr. A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature* **426**, 460–465 (2003).
24. Pomerening, J. R., Kim, S. Y. & Ferrell, J. E. Jr. Systems-level dissection of the cell-cycle oscillator: bypassing positive feedback produces damped oscillations. *Cell* **122**, 565–578 (2005).
25. Lahav, G. *et al.* Dynamics of the p53-Mdm2 feedback loop in individual cells. *Nature Genet.* **36**, 147–150 (2004).
26. Hardin, P. E. The circadian timekeeping system of *Drosophila*. *Curr. Biol.* **15**, R714–R722 (2005).
27. Vilar, J. M., Kueh, H. Y., Barkai, N. & Leibler, S. Mechanisms of noise-resistance in genetic oscillators. *Proc. Natl Acad. Sci. USA* **99**, 5988–5992 (2002).
28. Keener, J. & Sneyd, J. *Mathematical Physiology* (Springer, New York, 1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank U. Alon, D. Dubnau, J. Dworkin, A. Eldar, J. Ferrell, R. Kishony, B. Lazazzera, R. Losick, A. Raj, B. Shraiman, D. Sprinzak, M. Surette and members of the laboratory for comments. G.M.S. is supported by the Caltech Center for Biological Circuit Design. J.G.-O. acknowledges financial support from the Generalitat de Catalunya and the Ministerio de Educacion y Ciencia (Spain). M.B.E. acknowledges support from the Searle Scholars Program and the Burroughs Wellcome Fund CASI program.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.B.E. (melowitz@caltech.edu).

CHIP-mediated stress recovery by sequential ubiquitination of substrates and Hsp70

Shu-Bing Qian¹, Holly McDonough¹, Frank Boellmann¹, Douglas M. Cyr² & Cam Patterson¹

Exposure of cells to various stresses often leads to the induction of a group of proteins called heat shock proteins (HSPs, molecular chaperones)^{1,2}. Hsp70 is one of the most highly inducible molecular chaperones, but its expression must be maintained at low levels under physiological conditions to permit constitutive cellular activities to proceed^{3,4}. Heat shock transcription factor 1 (HSF1) is the transcriptional regulator of HSP gene expression⁵, but it remains poorly understood how newly synthesized HSPs return to basal levels when HSF1 activity is attenuated. CHIP (carboxy terminus of Hsp70-binding protein), a dual-function co-chaperone/ubiquitin ligase, targets a broad range of chaperone substrates for proteasomal degradation^{6–11}. Here we show that CHIP not only enhances Hsp70 induction during acute stress but also mediates its turnover during the stress recovery process. Central to this dual-phase regulation is its substrate dependence: CHIP preferentially ubiquitinates chaperone-bound substrates, whereas degradation of Hsp70 by CHIP-dependent targeting to the ubiquitin–proteasome system occurs when misfolded substrates have been depleted. The sequential catalysis of the CHIP-associated chaperone adaptor and its bound substrate provides an elegant mechanism for maintaining homeostasis by tuning chaperone levels appropriately to reflect the status of protein folding within the cytoplasm.

The activation status of HSF1 is governed in large part by the balance between the amount of misfolded proteins and the chaperone availability in the cell^{12,13}. However, HSF1 also seems to have additional layers of regulation in the cell¹⁴. For example, our previous data indicate that CHIP elicits the transcriptional activation of HSF1, independently of its ability to facilitate the clearance of misfolded proteins in the cell¹⁵. To improve our understanding of the effects of CHIP on both HSF1 activity and chaperone availability, we manipulated CHIP expression levels in HEK-293 cells by plasmid-mediated overexpression or double-stranded RNA (dsRNA)-dependent knockdown. In concordance with our previous observations, CHIP overexpression triggered the activation of HSF1 in a dose-dependent manner as indicated by luciferase reporter assays (Fig. 1a). Immunoblotting demonstrated elevated levels of Hsp70 as expected. Conversely, cells depleted of CHIP by dsRNA had lower basal levels of HSF1 activity. To our surprise, however, no decrease in Hsp70 levels was observed in these cells, and instead we found a moderate but consistent induction of Hsp70 protein levels (Fig. 1b). To explore this observation further, we analysed Hsp70 expression in *CHIP*^{−/−} and wild-type fibroblasts. Consistent with their reduced stress response in the absence of CHIP was our observation that *CHIP*^{−/−} cells demonstrated much lower Hsp70 induction after heat shock (42 °C, 15 min; Fig. 1c). However, the basal Hsp70 level before heat shock was paradoxically higher in *CHIP*^{−/−} than in *CHIP*^{+/+} cells. Taken together, these experiments imply that CHIP regulates Hsp70

expression through mechanisms that operate differentially under physiological and stressful conditions.

To assess the effects of CHIP on Hsp70 availability under conditions in which its ability to activate HSF1 is inoperative, we used an immortalized fibroblast cell line derived from *HSF1*^{−/−} mice. Recombinant adenovirus (AdV)-mediated CHIP overexpression markedly decreased Hsp70 in these cells, whereas no such effect was observed by overexpressing an inactive mutant of CHIP, CHIP(K30A), that does not dock with chaperones¹⁶ (Fig. 1d). Previous studies have indicated that *HSF1*^{−/−} cells are sensitive to stress because they lack an appropriate heat shock response¹⁷. We predicted that the diminished levels of Hsp70 after CHIP overexpression should render these cells more susceptible to stresses such as heat shock. As expected, *HSF1*^{−/−} cells were less viable when CHIP was overexpressed during incubation at 41 °C than were cells overexpressing CHIP(K30A) (Fig. 1e). Taken together, these observations indicate a role for CHIP in multiple steps of the stress response: while activating HSF1 and consequently inducing HSP expression, CHIP also negatively regulates steady-state levels of Hsp70 through an HSF1-independent mechanism.

As a co-chaperone ubiquitin ligase, CHIP is capable of ubiquitinating its associated chaperone partners *in vivo* and *in vitro*¹⁸, although it has not been apparent that this leads to their proteasomal degradation, nor have physiological roles for CHIP-dependent chaperone ubiquitination been assigned. To explore this relationship further, we examined turnover of HSPs in *CHIP*^{−/−} cells after reintroducing either CHIP, CHIP(K30A) or green fluorescent protein (GFP) alone as an additional control. Because the basal level of Hsp70 in mouse fibroblasts is very low, the infected cells were transiently exposed to 42 °C for 30 min and recovered at 37 °C for 6 h. The turnover of stress-induced Hsp70 was examined by incubating cells in the presence of the protein synthesis inhibitor cycloheximide. Hsp70 was remarkably stable ($t_{1/2} \gg 8$ h) in the absence of CHIP (Fig. 2a). However, adding back CHIP resulted in much shorter $t_{1/2}$ for Hsp70 (less than 4 h), whereas little change in $t_{1/2}$ was observed by adding back CHIP(K30A). In addition to Hsp70, CHIP interacts with Hsc70 and Hsp90, and CHIP also affected the stability of these proteins (Fig. 2a). After subtraction of their different turnover rates in the absence of CHIP, the most prominent net effect of CHIP was found on Hsp70 (Fig. 2a). The stability of Grp78, an endoplasmic reticulum-localized chaperone that does not interact with CHIP, was not affected by the manipulation of CHIP levels. The differential effects of CHIP on Hsc70 and Hsp70 were further confirmed in a pulse–chase analysis followed by immunoprecipitation (Fig. 2b).

We next examined whether the selective effects of CHIP on Hsp70 *in vivo* can be recapitulated *in vitro* with a reconstituted ubiquitination and degradation system. Equal amounts of Hsc70 and Hsp70

¹Carolina Cardiovascular Biology Center and ²Department of Cell and Developmental Biology, School of Medicine, University of North Carolina, Chapel Hill, North Carolina 27599, USA.

were incubated in a reaction containing E1, E2 (UbcH5a), CHIP, ubiquitin and ATP. Although CHIP is capable of targeting both chaperones for ubiquitination, Hsp70 was much more susceptible (Fig. 2c). The addition of purified 26S proteasomes resulted in rapid degradation of Hsp70, whereas only modest proteasome-dependent degradation was observed for Hsc70. No proteasomal degradation of Hsp70 was observed in the absence of CHIP (Fig. 2c, bottom panel). The degradative effects of CHIP on stress-inducible Hsp70 were further investigated by using ubiquitin mutants bearing a single lysine. We found that CHIP was able to initiate both canonical and non-canonical ubiquitin chains on both Hsc70 and Hsp70. However, Lys 48-linked high-molecular-weight species were selectively observed in reactions containing Hsp70 but not in those containing Hsc70 (Fig. 2d), which probably accounts for their differential susceptibility to proteasome-mediated degradation. Because Hsc70 and Hsp70 have nearly identical amino-acid sequences except for residues near the C-terminal region that are responsible for CHIP binding, it is possible that divergent residues flanking the binding sequence EEVD contribute partly to the differential activity of CHIP. Consistent with this notion is the observation that *in vitro* binding assays have shown different binding affinities of Hsc70 and Hsp70 for CHIP¹⁶.

When engaged by CHIP, Hsp70 functions as an adaptor for CHIP to access chaperone clients. Thus, it is unlikely that CHIP could negatively regulate Hsp70 availability while simultaneously facilitating the degradation of chaperone clients. We therefore postulated that a competitive relationship might exist in CHIP-mediated ubiquitination between adaptor and substrate. To test this hypothesis *in vivo*, we first examined the turnover of Hsp70 in the enforced presence of unfoldable cytoplasmic protein. Because proteotoxic

insults such as heat shock have various cellular effects, we performed these experiments with cytoplasmically expressed bovine serum albumin (cBSA). This protein is incapable of acquiring its normal conformation in the reducing cytosolic environment¹⁹. Its degradation by the proteasome can be further facilitated by CHIP overexpression (Supplementary Fig. 1). We found that expression of cBSA significantly decreased the proteasome-dependent turnover of Hsp70 (as demonstrated through proteasome inhibition with MG132) in comparison with the expression of β -galactosidase (β -Gal), a stably folded protein that is not a CHIP substrate (Fig. 3a).

The data in Fig. 3a imply that misfolded proteins compete with Hsp70 for the ubiquitin ligase activity of CHIP. However, there is no mutual competition between Hsp70 turnover and substrate degradation because Hsp70 overexpression does not interfere with the degradation of cBSA (Supplementary Fig. 2). This indicates that CHIP might preferentially ubiquitinate chaperone-bound substrates, whereas chaperone ubiquitination occurs after substrate abundance is depleted. To test this hypothesis directly, we reconstituted CHIP-dependent degradation by means of the proteasome. Purified Hsp70, in the presence of native firefly luciferase or thermally denatured luciferase (a known CHIP substrate²⁰), was incubated with E1, E2(UbcH5a) and CHIP in reaction buffer containing ubiquitin, ATP and 26S proteasome. Non-denatured luciferase was minimally degraded even after incubation for 3 h, whereas Hsp70 was degraded with linear kinetics in the same sample (Fig. 3b). In contrast, heat-denatured luciferase underwent rapid degradation ($t_{1/2} \approx 30$ min). Remarkably, under these conditions the steady-state level of Hsp70 was stable until most denatured luciferase had been degraded. The biphasic kinetics of Hsp70 degradation under these conditions is markedly different from the kinetics in the absence of denatured

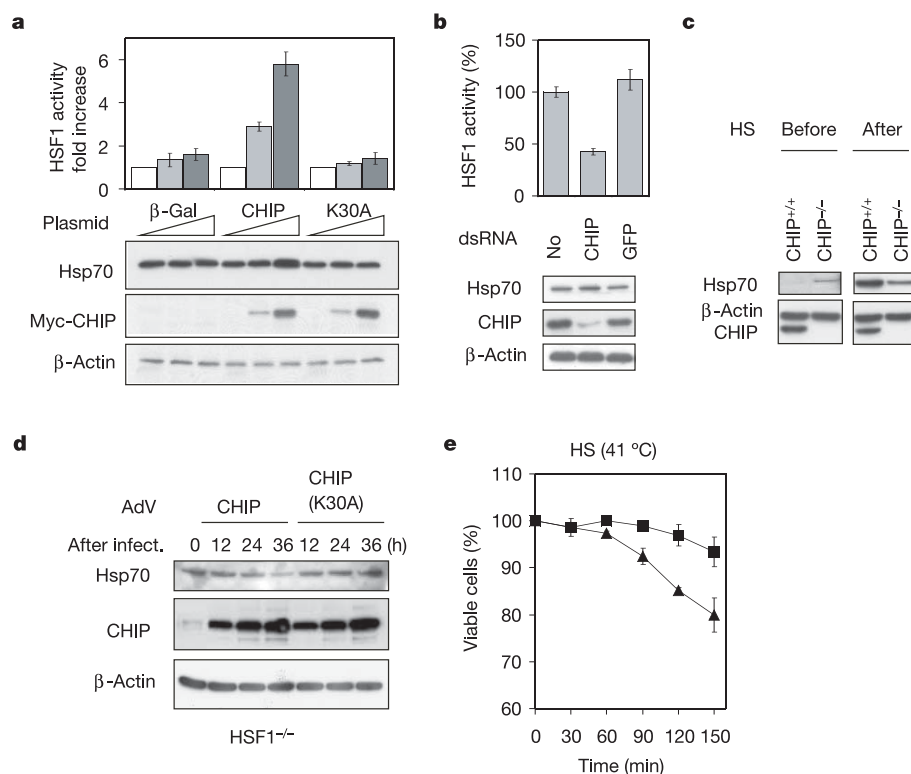


Figure 1 | CHIP regulates Hsp70 availability. **a**, HEK-293 cells were transfected with plasmids as indicated with 0 μ g (white bars), 0.1 μ g (grey bars) and 0.5 μ g (black bars) per well in a six-well plate. HSF1 activity was measured by dual luciferase reporter assay. Hsp70 levels were determined by immunoblotting. **b**, HEK-293 cells were transfected with dsRNA as indicated. Both HSF1 activity and Hsp70 levels were determined. **c**, Immortalized CHIP^{+/+} and CHIP^{-/-} cell lines were heat shocked for

15 min at 42 °C (HS) and left to recover for 4 h at 37 °C. Cell lysates from before and after heat shock were immunoblotted with Hsp70 antibody. **d**, HSF1^{-/-} cells were infected with AdV expressing CHIP or CHIP(K30A). Hsp70 levels were determined at different times after infection by immunoblotting of whole cell lysates. **e**, AdV-infected HSF1^{-/-} cells were incubated at 41 °C for various durations and cell viability was measured. Triangles, AdV/CHIP; squares, AdV/CHIP(K30A). Values are means $\pm$ s.d.

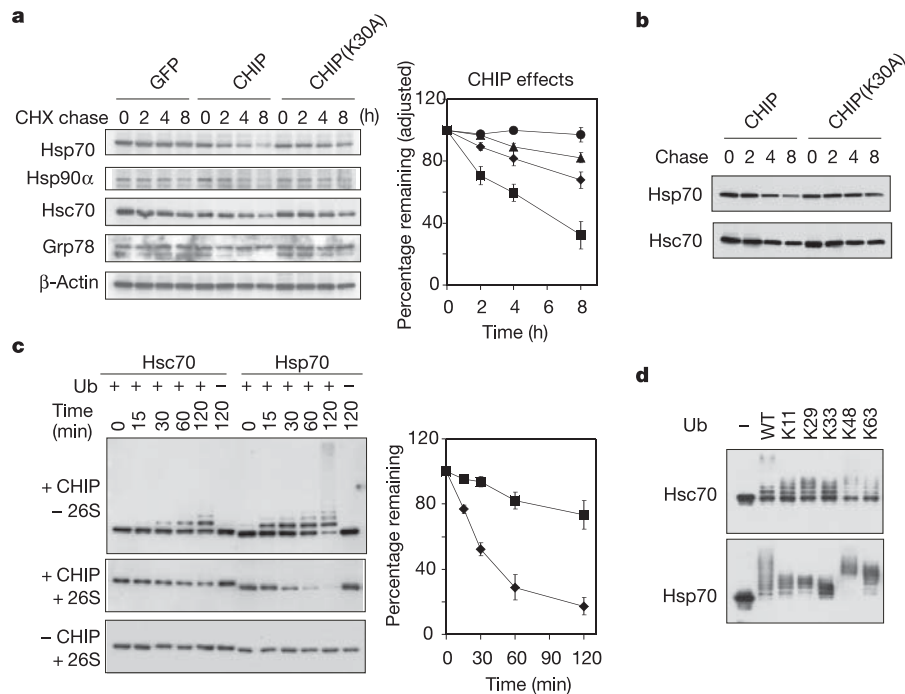


Figure 2 | Selective effects of CHIP on Hsp70 *in vivo* and *in vitro*. **a**, *CHIP*^{-/-} cells were infected with AdV expressing GFP, CHIP or CHIP(K30A). Infected cells were incubated for 30 min at 42 °C and left to recover for 6 h at 37 °C to induce Hsp expression. Cycloheximide (CHX) was then added and cell aliquots were collected at the times indicated. Whole cell lysates were immunoblotted with a panel of antibodies. The effects of CHIP on Hsp levels were quantified by subtracting their turnover in the presence of GFP. Squares, Hsp70; diamonds, Hsp90; triangles, Hsc70; circles, Grp78. Values are means ± s.d. **b**, AdV-infected *CHIP*^{-/-} cells were radiolabelled with [³⁵S]methionine for 30 min at 3 h after heat shock and chased with

unlabelled methionine for up to 8 h. Immunoprecipitations were performed with monoclonal antibody against Hsp70 or Hsc70. **c**, *In vitro* ubiquitination assays were performed by incubating equal amounts of Hsc70 and Hsp70 in reactions containing E1, E2 (UbcH5a), CHIP, ubiquitin (Ub) and ATP. Purified 26S proteasomes were included in the reaction mixture for the coupled degradation assay. Levels of Hsc70 (squares) and Hsp70 (diamonds) were determined by immunoblotting and quantified by densitometry. Values are means ± s.d. **d**, *In vitro* ubiquitination assays were performed with ubiquitin mutants bearing a single lysine as indicated, and the reaction was stopped after 120 min incubation. WT, wild-type.

substrate, supporting the concept of a hierarchy in the CHIP-targeted degradation of its chaperone adaptors and their associated substrates. On the basis of these observations, we further predicted that

stress-induced Hsp70 undergoes CHIP-mediated autoregulation during stress recovery in which most misfolded proteins are degraded or refolded²¹. To examine the physiological kinetics of stress-induced Hsp70 expression *in vivo*, we measured total Hsp70 levels in normal

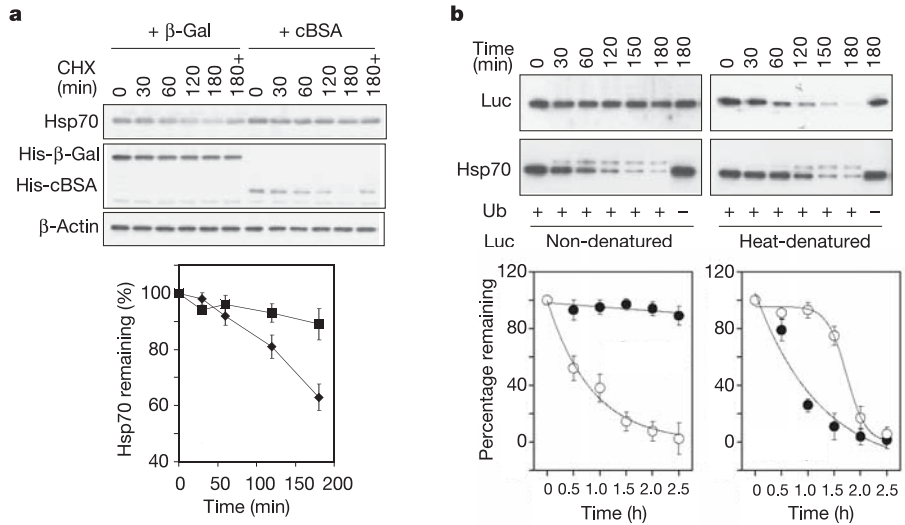


Figure 3 | Substrate-regulated Hsp70 turnover *in vivo* and *in vitro*. **a**, HEK-293 cells were co-transfected with plasmids expressing CHIP, Flag-Hsp70 and cBSA (squares) or β-Gal (diamonds). The turnover of Hsp70 was determined with the use of a cycloheximide (CHX) chase followed by immunoblotting. '180 + ' indicates the presence of 20 μM MG132 during the chase. Values are means ± s.d. **b**, *In vitro* degradation

assays were performed with Hsp70 (open circles) and luciferase (Luc; filled circles) as substrates. For heat-denatured luciferase, luciferase was heated at 43 °C for 10 min with Hsp70 and Hdj2 (ref. 20). Both the luciferase and Hsp70 levels were determined in the same samples by immunoblotting and quantified by densitometry. Values are means ± s.d.

mouse fibroblasts during the acute phase of the stress response and the chronic phase of stress recovery. Hsp70 was rapidly induced to its peak level about 4 h after heat shock (42 °C, 10 min) and maintained its steady-state level for about 8 h, followed by a decrease (Fig. 4a). This biphasic pattern echoes the substrate-dependent Hsp70 turnover observed in reconstitution assays *in vitro* (Fig. 3b). The degradation of misfolded proteins approximates first-order kinetics (Fig. 3a), whereas the stability of Hsp70 is maintained until substrate concentrations fall below a threshold, at which time Hsp70 degradation accelerates rapidly (Figs 3b and 4a). Further supporting the model that threshold levels of misfolded substrates determine when the clearance of excess Hsp70 begins, there is a close correlation between the intensities of applied stress and the sustained period of Hsp70 expression (Supplementary Fig. 3).

The results presented here show a previously unknown mechanism to regulate inducible HSP expression and resolution of the stress response. Our observations indicate that the effect of CHIP on Hsp70 might not counteract its active role in the clearance of misfolded proteins and that both functions might instead be integrally related during the stress response and recovery process. To define the role of CHIP in orchestrating both phases, we examined the behaviour of Hsp70 in *CHIP*^{-/-} cells during acute stress and the chronic recovery phase before and after reintroducing CHIP with the use of AdV. As expected from previous observations, adding back CHIP, but not CHIP(K30A), resulted in enhanced Hsp70 induction after heat shock, which was consistent with the role of CHIP in activating HSF1 (Fig. 4b). During the recovery phase, however, a rapid decrease in steady-state levels of Hsp70 was observed only in the presence of wild-type CHIP. In contrast, cells infected with AdV expressing GFP or CHIP(K30A) showed a prolonged plateau of

Hsp70 levels. The role of CHIP in the regulation of Hsp70 availability was further confirmed by pulse-chase analysis followed by immunoprecipitation, which indicated that the *t*_{1/2} of Hsp70 was much shorter during the stress recovery phase when CHIP was reintroduced into *CHIP*^{-/-} cells (Fig. 4c). Thus, the presence of CHIP actively maintains cellular homeostasis not only by enhancing Hsp70 induction during the acute phase of stress but also by removing excess Hsp70 during the recovery process.

The U-box is required for CHIP ubiquitin ligase activity but is dispensable in the stimulation of HSF1 activity¹⁵. We therefore predicted that inactivation of the CHIP U-box would lead to sustained high levels of Hsp70 in the cell after heat shock. CHIP(H260Q), a U-box mutant that does not bind E2, shows dominant-negative effects in HEK-293 cells. Indeed, we observed a much higher basal level of Hsp70 in cells transfected with CHIP(H260Q) (Fig. 4d). After a mild heat shock, the amount of Hsp70 was largely sustained in cells expressing CHIP(H260Q), whereas overexpression of CHIP accelerated both the elevation and restoration of Hsp70 levels (Fig. 4e). These results confirm that the ubiquitin ligase activity of CHIP is required for Hsp70 degradation during stress recovery *in vivo*.

Several principles emerge from our results. First, CHIP-mediated Hsp70 turnover represents a novel regulatory mechanism for Hsp70 in addition to HSF1-mediated transcriptional regulation. Second, these data indicate that there are multiple coordinated steps by which the chaperone and ubiquitin-proteasome systems interact to regulate the stress response. Without appropriate CHIP function, there is impaired ability to induce Hsp70 during stress and concomitantly a deficient buffering of misfolded proteins. At the same time, CHIP is required for proteasome-dependent clearance of unfoldable proteins

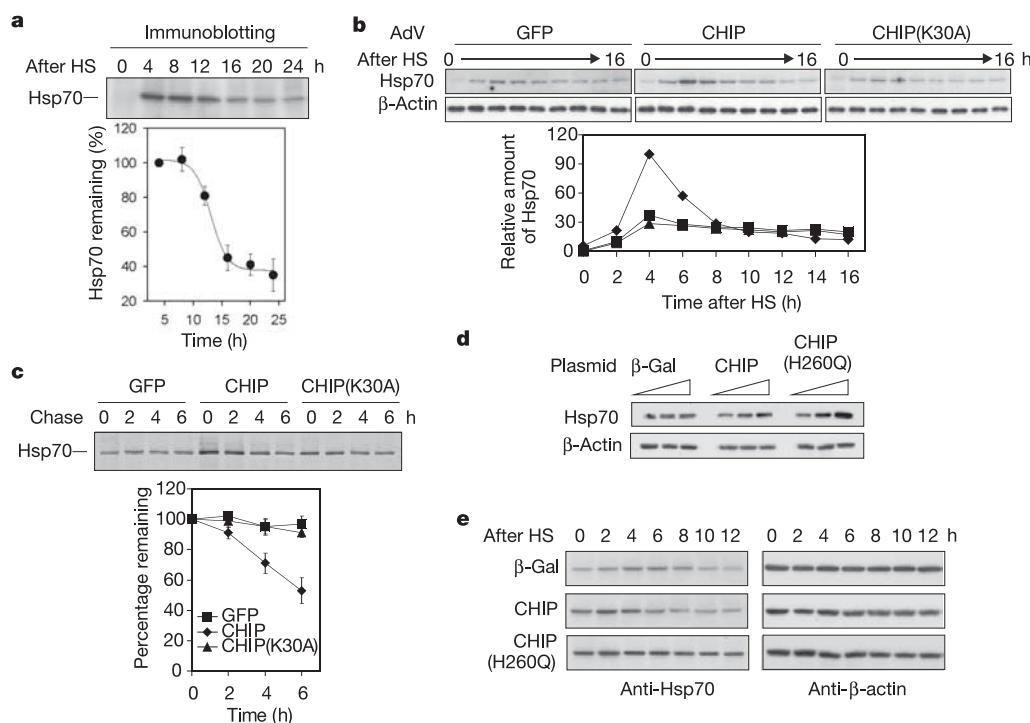


Figure 4 | CHIP orchestrates both stress response and recovery processes. **a**, Mouse fibroblasts were heat shocked (HS) at 42 °C for 10 min, and aliquots were collected every 4 h during recovery at 37 °C. Hsp70 levels were determined by immunoblotting. Values are means $\pm$ s.d. **b**, *CHIP*^{-/-} cells were infected with AdV expressing GFP (squares), CHIP (diamonds) or CHIP(K30A) (triangles). Infected cells were heat shocked at 42 °C for 10 min, and aliquots were collected every 2 h during recovery at 37 °C. Hsp70 levels were determined by immunoblotting and quantified by densitometry. **c**, The same AdV-infected *CHIP*^{-/-} cells as in **b** were heat

shocked at 42 °C for 10 min, and radiolabelled with [³⁵S]methionine for 30 min at 3 h after HS and chased with unlabelled methionine for up to 6 h. Immunoprecipitation was performed with a Hsp70-specific antibody and quantified with a PhosphorImager. Symbols are as in **b**. Values are means $\pm$ s.d. **d**, HEK-293 cells were transfected with plasmids as indicated with 0, 0.1 and 0.5 μ g per well in a six-well plate. Hsp70 levels were determined by immunoblotting. **e**, Transfected HEK-293 cells were heat shocked 42 °C for 5 min, and aliquots were collected every 2 h during recovery at 37 °C. Hsp70 levels were determined by immunoblotting.

and the recovery from the stress response by means of Hsp70 degradation. Finally, CHIP-mediated sequential catalysis of the chaperone adaptor and substrate indicates an elegant and simple mechanism to regulate the target specificity for a single ubiquitin ligase. In the multisubunit ubiquitin ligases such as Skp1-Cdc53/Cul1-F-box (SCF), the F-box adaptor targets a distinct array of substrates for ubiquitination and proteasomal degradation²². Intriguingly, most F-box adaptors in the SCF complex are labile proteins as a result of autocatalysis by the intrinsic E3 activity^{23–25}. In the case of CHIP, Hsp70 acts as an adaptor, or 'F-box equivalent', for CHIP to ubiquitinate chaperone clients; degradation of Hsp70 may therefore be a substrate-dependent variation on this autocatalytic mechanism.

METHODS

Cell lines, plasmids and viruses. HEK-293 cells were maintained in DMEM medium containing 10% FBS. *CHIP*^{+/+} and *CHIP*^{-/-} cell lines were established from fibroblasts isolated from adult lungs of *CHIP*^{+/+} and *CHIP*^{-/-} mice by transfection with SV40 large T antigen (kindly provided by D. Ron). The *HSF1*^{-/-} murine embryonic fibroblast cell line was provided by I. Benjamin. Mammalian expression vectors for β -Gal, cBSA, Hsp70, CHIP and CHIP(K30A) are based on pcDNA3.1. Recombinant adenoviruses expressing CHIP, CHIP(K30A) or GFP have been described previously¹⁵.

Antibodies and short interfering RNAs. Monoclonal anti-Hsp70 (SPA810), anti-Hsp90 α (SPA845) and anti-Hsc70 (SPA815) were purchased from Stressgen. Monoclonal anti-Grp78 antibody was from Santa Cruz. Monoclonal anti-His antibody was from Qiagen. Monoclonal anti- β -actin was from Sigma. Polyclonal anti-luciferase was from Promega. Double-stranded RNA for CHIP (target sequence 5'-GGAGCAGGCGCAAUCGUCUG-3') and GFP (target sequence 5'-AACGAGAAGCGCAUCACAUG-3') were synthesized by Dharmacon.

Dual luciferase assay. A firefly luciferase gene under the control of the strictly stress-regulated human *Hsp70B* promoter was used to assess HSF1 transcriptional activity. To control for the differences in transfection efficiency and other effects, a *Renilla* luciferase gene driven by a cytomegalovirus promoter was included in all transfections²⁶. Reporter activity was determined with a dual-luciferase reporter assay (Promega).

In vitro reconstitution assay. His-tagged CHIP, and UbcH5a, were produced in *Escherichia coli* BL21(DE3) followed by purification with Ni²⁺-nitrilotriacetate agarose. *In vitro* ubiquitination assays were performed in the presence of 0.1 μ M purified rabbit E1 (BioMol), 2 μ M UbcH5a, 3 μ M CHIP, 50 μ M ubiquitin (Sigma), 1 mM dithiothreitol, 2 mM MgCl₂ and 4 mM ATP. For *in vitro* degradation assays, 50 nM purified 26S proteasomes (BioMol) was included in the reaction mixture. Reactions were performed at 37°C, and samples were analysed by SDS-PAGE and immunoblotting with appropriate antibodies. Heat treatment of firefly luciferase was described elsewhere²¹. In brief, 0.5 μ M luciferase was heated at 43°C for 10 min with 1 μ M Hsp70 and 1 μ M Hdj2 on the presence of 4 mM ATP in 50 mM Tris-HCl pH 7.5 containing 2 mM MgCl₂. After being heated, samples were quickly chilled in an ice bath. Aliquots were incubated in the reaction mixture as indicated above.

Pulse-chase analysis. Cells were radiolabelled with [³⁵S]Met (100 μ Ci) for the indicated durations. After washing with PBS containing excess unlabelled Met (1 mg ml⁻¹), cells were chased in the complete DMEM for the indicated duration. Whole cell lysates were made in Tris-buffered saline pH 7.5 containing 1% Triton X-100 and 10 U ml⁻¹ DNase A (Roche). Lysates were either directly resolved by SDS-PAGE or used for immunoprecipitation with Protein G beads coated in monoclonal anti-Hsp70 (SPA810).

Received 14 December 2005; accepted 25 January 2006.

- Parsell, D. A. & Lindquist, S. The function of heat-shock proteins in stress tolerance: degradation and reactivation of damaged proteins. *Annu. Rev. Genet.* **27**, 437–496 (1993).
- Morimoto, R. I. Cells in stress: transcriptional activation of heat shock genes. *Science* **259**, 1409–1410 (1993).
- Feder, J. H., Rossi, J. M., Solomon, J., Solomon, N. & Lindquist, S. The consequences of expressing hsp70 in *Drosophila* cells at normal temperatures. *Genes Dev.* **6**, 1402–1413 (1992).

- Volloch, V. Z. & Sherman, M. Y. Oncogenic potential of Hsp72. *Oncogene* **18**, 3648–3651 (1999).
- Wu, C. Heat shock transcription factors: structure and regulation. *Annu. Rev. Cell Dev. Biol.* **11**, 441–469 (1995).
- Cyr, D. M., Hohfeld, J. & Patterson, C. Protein quality control: U-box-containing E3 ubiquitin ligases join the fold. *Trends Biochem. Sci.* **27**, 368–375 (2002).
- Connell, P. et al. The co-chaperone CHIP regulates protein triage decisions mediated by heat-shock proteins. *Nature Cell Biol.* **3**, 93–96 (2001).
- Meacham, G. C., Patterson, C., Zhang, W., Younger, J. M. & Cyr, D. M. The Hsc70 co-chaperone CHIP targets immature CFTR for proteasomal degradation. *Nature Cell Biol.* **3**, 100–105 (2001).
- Demand, J., Alberti, S., Patterson, C. & Hohfeld, J. Cooperation of a ubiquitin domain protein and an E3 ubiquitin ligase during chaperone/proteasome coupling. *Curr. Biol.* **11**, 1569–1577 (2001).
- Xu, W. et al. Chaperone-dependent E3 ubiquitin ligase CHIP mediates a degradative pathway for c-ErbB2/Neu. *Proc. Natl Acad. Sci. USA* **99**, 12847–12852 (2002).
- McDonough, H. & Patterson, C. CHIP: a link between the chaperone and proteasome systems. *Cell Stress Chaperones* **8**, 303–308 (2003).
- Shi, Y., Mosser, D. D. & Morimoto, R. I. Molecular chaperones as HSF1-specific transcriptional repressors. *Genes Dev.* **12**, 654–666 (1998).
- Zou, J., Guo, Y., Guettouche, T., Smith, D. F. & Voellmy, R. Repression of heat shock transcription factor HSF1 activation by HSP90 (HSP90 complex) that forms a stress-sensitive complex with HSF1. *Cell* **94**, 471–480 (1998).
- Ahn, S. G. & Thiele, D. J. Redox regulation of mammalian heat shock factor 1 is essential for Hsp gene activation and protection from stress. *Genes Dev.* **17**, 516–528 (2003).
- Dai, Q. et al. CHIP activates HSF1 and confers protection against apoptosis and cellular stress. *EMBO J.* **22**, 5446–5458 (2003).
- Ballinger, C. A. et al. Identification of CHIP, a novel tetratricopeptide repeat-containing protein that interacts with heat shock proteins and negatively regulates chaperone functions. *Mol. Cell. Biol.* **19**, 4535–4545 (1999).
- McMillan, D. R., Xiao, X., Shao, L., Graves, K. & Benjamin, I. J. Targeted disruption of heat shock transcription factor 1 abolishes thermotolerance and protection against heat-inducible apoptosis. *J. Biol. Chem.* **273**, 7523–7528 (1998).
- Jiang, J. et al. CHIP is a U-box-dependent E3 ubiquitin ligase: identification of Hsc70 as a target for ubiquitylation. *J. Biol. Chem.* **276**, 42938–42944 (2001).
- Guo, Y. et al. Evidence for a mechanism of repression of heat shock factor 1 transcriptional activity by a multichaperone complex. *J. Biol. Chem.* **276**, 45791–45799 (2001).
- Murata, S., Minami, Y., Minami, M., Chiba, T. & Tanaka, K. CHIP is a chaperone-dependent E3 ligase that ubiquitylates unfolded protein. *EMBO Rep.* **2**, 1133–1138 (2001).
- Weibezahn, J. et al. Thermotolerance requires refolding of aggregated proteins by substrate translocation through the central pore of ClpB. *Cell* **119**, 653–665 (2004).
- Petroski, M. D. & Deshaies, R. J. Function and regulation of cullin-RING ubiquitin ligases. *Nature Rev. Mol. Cell Biol.* **6**, 9–20 (2005).
- Galan, J. M. & Peter, M. Ubiquitin-dependent degradation of multiple F-box proteins by an autocatalytic mechanism. *Proc. Natl Acad. Sci. USA* **96**, 9124–9129 (1999).
- Zhou, P. & Howley, P. M. Ubiquitination and degradation of the substrate recognition subunits of SCF ubiquitin-protein ligases. *Mol. Cell* **2**, 571–580 (1998).
- Wee, S., Geyer, R. K., Toda, T. & Wolf, D. A. CSN facilitates Cullin-RING ubiquitin ligase function by counteracting autocatalytic adapter instability. *Nature Cell Biol.* **7**, 387–391 (2005).
- Boellmann, F. et al. DAXX interacts with heat shock factor 1 during stress activation and enhances its transcriptional activity. *Proc. Natl Acad. Sci. USA* **101**, 4100–4105 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Thiele for critical reading of the manuscript. This work was supported by grants from the National Institutes of Health, the American Heart Association and the Burroughs Wellcome Fund.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.P. (cpatters@med.unc.edu).

LETTERS

RNA-mediated response to heat shock in mammalian cells

Ilya Shamovsky^{1,2}, Maxim Ivannikov¹, Eugene S. Kandel³, David Gershon^{2,4} & Evgeny Nudler¹

The heat-shock transcription factor 1 (HSF1) has an important role in the heat-shock response in vertebrates by inducing the expression of heat-shock proteins (HSPs) and other cytoprotective proteins¹. HSF1 is present in unstressed cells in an inactive monomeric form and becomes activated by heat and other stress stimuli. HSF1 activation involves trimerization and acquisition of a site-specific DNA-binding activity^{2,3}, which is negatively regulated by interaction with certain HSPs^{4–6}. Here we show that HSF1 activation by heat shock is an active process that is mediated by a ribonucleoprotein complex containing translation elongation factor eEF1A and a previously unknown non-coding RNA that we term HSR1 (heat shock RNA-1). HSR1 is constitutively expressed in human and rodent cells and its homologues are functionally interchangeable. Both HSR1 and eEF1A are required for HSF1 activation *in vitro*; antisense oligonucleotides or short interfering (si)RNA against HSR1 impair the heat-shock response *in vivo*, rendering cells thermosensitive. The central role of HSR1 during heat shock implies that targeting this RNA could serve as a new therapeutic model for cancer, inflammation and other conditions associated with HSF1 deregulation.

HSF1 activation has been proposed to occur spontaneously through relief of negative regulation imposed by chaperones^{6–8}. Activated HSF1 binds to the heat shock element (HSE), a consensus sequence in HSP promoters composed of inverted NGAAN repeats⁹, and rescues the RNA polymerase II elongation complex from promoter-proximal arrest^{10,11}. A large portion of the inactive HSF1 in unstressed cells is believed to be complexed with HSP90, P23 and an immunophilin^{12–14}. HSP90 has emerged as a primary candidate for the role of repressor of both HSF oligomerization and transcriptional competence⁵. Some data, however, point to the existence of auxiliary cellular factors that regulate HSF1 activation. The temperature threshold of HSF1 activation *in vivo* during heterologous expression is reprogrammed to that of the host cell^{12,15} and is higher in motor neurons and lower in T lymphocytes than in other mammalian cells^{16,17}. Moreover, the fast HSF1 activation kinetics in response to heat shock¹⁸ is difficult to reconcile with a simple diffusion-controlled mechanism (Supplementary Information).

To identify putative auxiliary factors involved in HSF1 activation, we looked at proteins from whole-cell lysates retained on covalently immobilized HSF1. HSF1 was expressed in *E. coli* and purified as a glutathione S-transferase (GST) fusion protein. The purification procedure was optimized to minimize spontaneous trimerization of HSF1 (see Supplementary Information). A polypeptide of ~45 kDa was retained on HSF1-Sepharose after incubation with a lysate from heat-shocked, but not untreated, BHK-21 (Fig. 1a) or HeLa cells (Supplementary Fig. 1). The bound polypeptide was identified by MALDI-TOF (matrix-assisted laser desorption/ionization–time of flight) as translation elongation factor eEF1A. Retention of eEF1A was temperature-sensitive, as incubation at 43 °C released most of

the bound eEF1A (Fig. 1a). No eEF1A binding was observed if bovine serum albumin (BSA)-Sepharose was used in place of HSF1-Sepharose.

We next examined whether the eEF1A-containing fraction eluted by heat from HSF1-Sepharose affected HSF1 function *in vitro*. Notably, the eEF1A fraction activated endogenous HSF1 in the lysate of unstressed cells as shown by an electromobility shift assay (EMSA) (Fig. 1b). Furthermore, this fraction induced DNA-binding activity of recombinant HSF1 (Fig. 1c). The effect was dose-dependent with respect to both the amount of HSF1 (or cell lysate, Supplementary Fig. 1c) and eEF1A fraction (Fig. 1c). In our system, heating lysate from unstressed cells in the absence of the eEF1A fraction did not yield any detectable HSF1 activity (Fig. 1b, lane 1, see Supplementary Information). The induced HSF1 DNA-binding activity was specific, as it was sensitive to both an anti-HSF antibody and an excess of unlabelled HSE oligonucleotide (Fig. 1b, lanes 6, 7 and 8, 9). Activation of purified HSF1 by the eEF1A fraction was accompanied by trimerization of HSF as confirmed by protein–protein cross-linking (Fig. 1d).

Co-immunoprecipitation of eEF1A and HSF1 from a whole-cell lysate (Fig. 1e) revealed the presence of a small amount of HSF1–eEF1A complexes in unstressed cells (lanes 1–4) with the amount increasing during heat-shock treatment, reaching a plateau after 30 min of heat shock (lanes 6, 7) and returning to the initial level after a 1-h recovery at 37 °C (lane 8). These results, together with elution of eEF1A from HSF1-Sepharose at 43 °C (Fig. 1a), indicate that eEF1A has a higher affinity for the HSF1 trimer than the monomer.

Release of HSF1 from the inhibitory multichaperone complex during heat shock⁴ could also contribute to the greater availability of HSF1 for eEF1A. Moreover, free eEF1A should accumulate in the cell under heat-shock conditions owing to general translational shut-down¹⁹ and cytoskeleton collapse²⁰. These events would ensure the availability of eEF1A molecules capable of forming a complex with HSF1, and thus link the major cellular perturbations caused by heat shock with HSF1 activation.

Our initial attempts to activate HSF1 with eEF1A isolated from rat liver or HeLa cells were unsuccessful, suggesting that the eEF1A fraction eluted from HSF1-Sepharose contained unidentified component(s) required for HSF1 activation. As eEF1A normally binds aminoacyl-tRNA, we sought to determine whether this interaction has a role in HSF1 activation. As shown in Fig. 2a, HSF1 binding to DNA was strongly inhibited *in vitro* by pre-incubation of the cell lysate with RNase A. The effect was specific, as addition of excess tRNA before treatment with RNase protected HSF1 activity. Similar results were obtained using micrococcal nuclease in the presence of Ca²⁺ (not shown). These data prompted us to search for a specific RNA in the eEF1A fraction that might participate in HSF1 activation. This RNA, which we term HSR1, was detected as a single band migrating at ~2 kb on denaturing polyacrylamide gel electrophoresis

¹Department of Biochemistry, New York University School of Medicine, New York, New York 10016, USA. ²Faculty of Biology, Technion-Israel Institute of Technology, Haifa 32000, Israel. ³The Cleveland Clinic Foundation Lerner Research Institute, Cleveland, Ohio 44195, USA. ⁴Redox Pharmaceutical Corp, Greenvale, New York 11548, USA.

(PAGE) (Fig. 2b). The band was sensitive to RNase A, but not to DNase (Fig. 2b). Addition of purified HSR1 (but not total RNA) to a lysate of heat-shocked BHK cells treated sequentially with micrococcal nuclease (to remove endogenous HSR1) and EGTA (to inactivate the nuclease) restored the DNA-binding activity of HSF1 (Supplementary Fig. 2a).

We cloned HSR1 from BHK and HeLa cells, sequenced it, and found it to be a ~600-nucleotides (nt) long RNA without a poly(A) tail (see Supplementary Information). Sequence comparison of HSR1 from BHK and HeLa cells revealed a high degree of homology, with only a 4-nt difference. Northern blot analysis showed that HSR1 is constitutively expressed in BHK and HeLa cells, and its level seems unaffected by heat shock (Supplementary Fig. 2b).

For the reconstitution experiment shown in Fig. 2c, recombinant mouse HSF1 was used along with eEF1A purified from rat liver²¹. Both HSR1 isolated from heat-shocked BHK or HeLa cells (lanes 7, 8) and HSR-T3 (*in vitro* synthesized HSR1, lane 10) activated HSF1 when added together with purified eEF1A. Neither component alone was capable of activating HSF1 (lanes 2–6). The slight stimulatory effect of isolated eEF1A (lane 2) on HSF1 activation resulted from residual amounts of co-purified HSR1, and was eliminated by RNase A treatment (not shown). Crosslinking with ethylene glycol-bis-succinimidylsuccinate (EGS) confirmed that HSR-T3-mediated activation was accompanied by HSF1 trimerization (see Supplementary Fig. 2c). In contrast, HSR-T7 (antisense HSR) failed to induce HSF1 binding to DNA (Fig. 2c, lane 9) and suppressed the activating effect of HSR-T3 (lane 11). We conclude that HSR1 and eEF1A are necessary and sufficient for HSF1 activation *in vitro*.

Notably, the mobility of HSR-T3 depended on the conditions to which it was exposed before PAGE analysis. Both Mg²⁺ and elevated temperature caused *in vitro* synthesized HSR to migrate similarly to HSR1 isolated from heat-shocked cells (Fig. 2d, compare lanes 2 and 3, 4). This conformational change of HSR1 may be associated with triggering HSF1 activation, as only the slow-migrating form of HSR1

was retained on HSF1-Sepharose (Fig. 2b). Polymerase chain reaction (PCR) analysis of genomic DNA using HSR1-specific primers produced a single product of ~600 nt, ruling out RNA processing as the cause of variability in HSR1 mobility and indicating that the cloned HSR fragment indeed represents full-length HSR1.

To obtain evidence of *in vivo* complex formation between HSF1, eEF1A and HSR1, we extracted RNA from immunoprecipitates corresponding to lanes 5 and 7 of Fig. 1e and analysed them by PCR with reverse transcription (RT-PCR). A greater amount of HSR1 was precipitated from heat-shocked cells compared with the control cells (lanes 3 and 4, Fig. 2e), indicating that the amount of HSR1–eEF1A complexes *in vivo* increased upon heat shock.

To delineate essential functional domains within HSR1, we used a set of 15 overlapping 45-mer antisense DNA oligonucleotides covering the entire HSR1 sequence (Fig. 2f). Each oligonucleotide was tested in the reconstituted system for its ability to suppress HSF1 activation by HSR1–eEF1A. The results of this experiment identified two domains in HSR1 that were essential for HSF1 activation: four oligonucleotides—1^{HSR} and 2^{HSR}, which span 84 nt at the 5' terminus of HSR1; and 5^{HSR} and 6^{HSR}, which correspond to nucleotides 157–240—inhibited HSF1 activation by more than 90% (Fig. 2f).

We used the inhibitory oligonucleotide 6^{HSR} to suppress the heat-shock response *in vivo*. BHK cells were transfected with either 6^{HSR} or a control double-stranded (ds) oligonucleotide (6^{HSR}/anti-6^{HSR}), and subjected to heat-shock treatment. HSF1 activation was monitored by EMSA (Fig. 3a) and HSP72 synthesis was assessed by immunoblotting after recovery of the cells for 16 h at 37 °C for following heat shock (lower panel). We found that the 6^{HSR} oligonucleotide, but not the 6^{HSR}/anti-6^{HSR} ds-oligonucleotide, inhibited HSF1 activation *in vivo* in addition to the production of HSP72 in response to heat shock (Fig. 3a). Consistently, 6^{HSR} compromised the induction of thermotolerance, demonstrated by a reduction in cell survival following a lethal heat-shock challenge (Fig. 3b). Transfection with 6^{HSR} resulted in more than 80% lethality after 60 min of heat

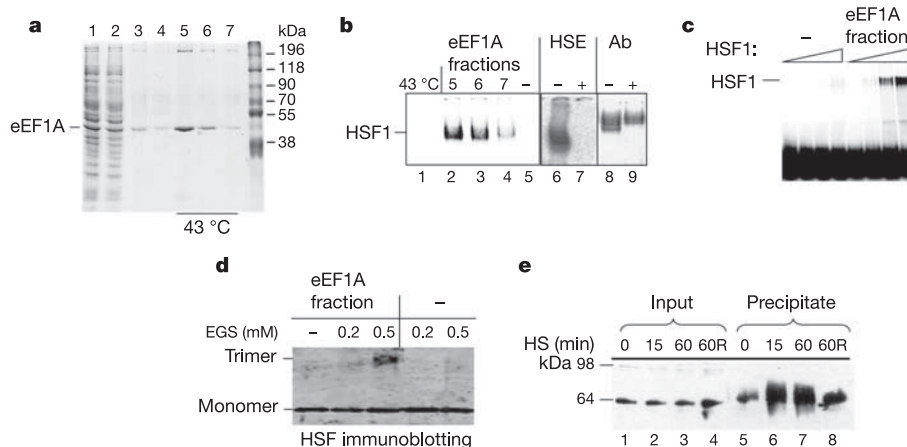


Figure 1 | Identification of an HSF1-activating fraction containing eEF1A. **a**, Fractionation of a lysate of heat-shocked BHK cells on HSF1-Sepharose. Lane 1, whole-cell lysate of heat-shocked BHK cells; lane 2, supernatant after incubation of BHK lysate with HSF1 Sepharose; lane 3, HSF1-Sepharose beads incubated with the lysate and washed; lanes 4, HSF1-Sepharose beads after three successive rounds of elution at 43 °C; lanes 5–7, proteins released from HSF1-Sepharose after successive elutions at 43 °C. **b**, The eEF1A-containing fraction activates HSF1 in a lysate from unstressed cells. An aliquot of a whole-cell lysate of unstressed BHK cells (20 µg) was incubated in the absence (lanes 1, 5) or in the presence (lanes 2–4) of fractions 5–7 from Fig. 1a and analysed by EMSA. '43 °C' indicates that the lysate was heat-shocked for 15 min before EMSA (lane 1). Lanes 6–9 contain fraction 5. 'HSE' (lanes 6, 7) indicates the inclusion of an excess of unlabelled HSE oligonucleotide. 'Ab' indicates the presence of a monoclonal anti-HSF1

antibody that causes a supershift (lane 9). **c**, *In vitro* activation of recombinant HSF1 by the eEF1A fraction. Increasing amounts of purified recombinant mouse HSF1 were incubated in the absence (–) or presence of the eEF1A fraction and analysed by EMSA. **d**, The eEF1A fraction induces trimerization of purified recombinant HSF1. Purified recombinant mouse HSF1 was incubated in the absence (–) or presence of the eEF1A-containing fraction and the crosslinking reagent EGS. Proteins were then separated by SDS-PAGE, transferred to a nitrocellulose membrane, and probed with an anti-HSF1 antibody. **e**, Formation of an eEF1A–HSF1 complex *in vivo*. BHK cells were heat-shocked (HS) for the times indicated and whole-cell lysates were analysed by immunoprecipitation with an anti-eEF1A antibody followed by immunoblotting with anti-HSF1. '60R' indicates 60-min recovery after 60-min heat shock.

shock at 45 °C, whereas the control anti- 6^{HSR} oligonucleotide had no significant effect on cell viability (Fig. 3b). Similar results were obtained with HeLa cells (not shown).

We obtained independent evidence for an essential role of HSR1 in the heat-shock response *in vivo* from RNA interference (RNAi) experiments. On the basis of data obtained using antisense oligonucleotides, we constructed vectors expressing siRNA against different domains of HSR1 (siRNA^{HSR}). The effect of siRNA was monitored by the heat-shock-induced expression of a plasmid-derived *Renilla* luciferase (RLuc) reporter fused to an inducible human *HSP70* promoter (see Methods). We co-transfected this construct into HeLa cells together with RNAi vectors and measured changes in RLuc activity induced by heat shock. Whereas RLuc activity was induced ~200-fold by heat-shock treatment followed by recovery at 37 °C, the siRNA^{HSR} corresponding to the 6^{HSR} antisense oligonucleotide strongly inhibited heat-shock induction of RLuc (Fig. 3c). Notably, a mutant construct carrying a single G → C substitution in the siRNA sequence had a significantly smaller effect on RLuc induction by heat shock.

Finally, we generated HeLa cell lines that stably expressed siRNA^{HSR}. These cells were transiently transfected with the RLuc reporter plasmid, and heat-shock induction of RLuc activity was monitored. We found that cells expressing siRNA^{HSR} or full-length antisense

HSR1, but not green fluorescent protein (GFP), were deficient in their ability to induce RLuc activity after a 2-h heat shock at 43 °C (Fig. 3d). Consistently, heat-shock induction of HSF1 DNA-binding activity was severely impaired in siRNA^{HSR}, but not in GFP-expressing cells (Fig. 3d, inset). Predictably, cells stably expressing siRNA^{HSR} failed to acquire thermotolerance after heat-shock pre-conditioning (Fig. 3e). Taken together, these data show that HSR1 is essential for the heat-shock response in mammalian cells.

It has been argued that the inhibitory complex of HSF1 with HSP90 and other chaperones dissociates as a result of competition between denatured cellular proteins and HSF for binding to HSP90^{4,12,13,22}. Here we demonstrate that HSF1 activation during heat shock is enhanced and perhaps triggered by specific cellular factors, including translation elongation factor eEF1A and a previously unknown untranslated RNA, HSR1 (Supplementary Fig. 4). HSF1 activation *in vitro* can be achieved in a reconstituted system with eEF1A+HSR1. Although purified HSF1 has been reported to trimerize spontaneously *in vitro* in response to elevated temperature, pH and other stimuli^{23–25}, our data suggest that the concentration threshold at which spontaneous HSF activation occurs²⁶ *in vitro* exceeds the effective intracellular HSF1 concentration, particularly during the first moments of heat-shock treatment. Nonetheless, the

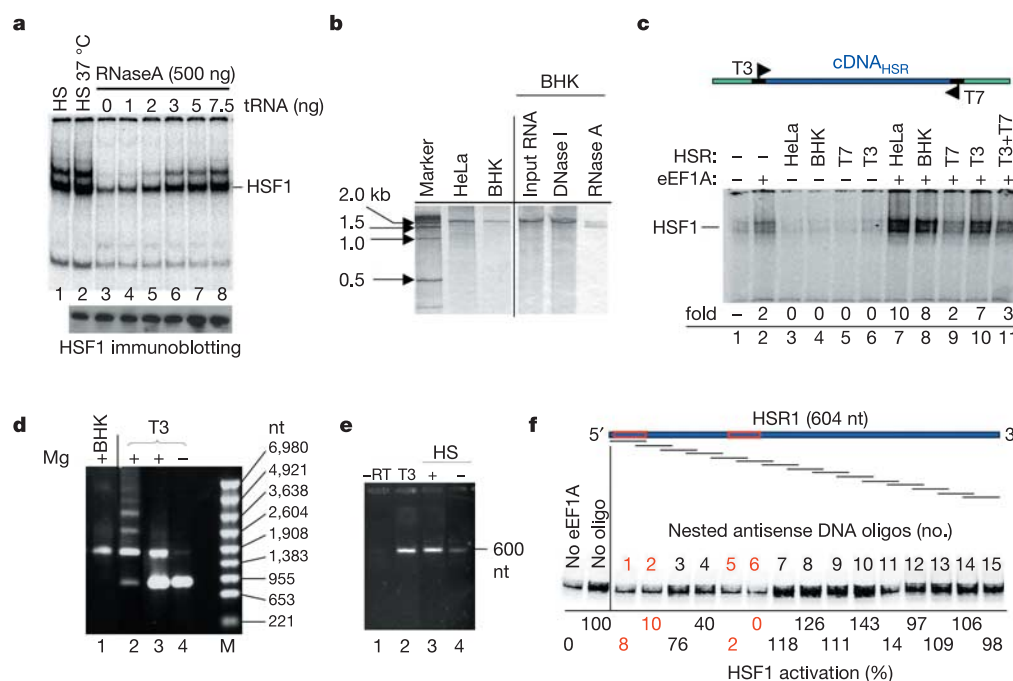


Figure 2 | HSR1-mediated activation of HSF1. **a**, Inhibition of HSF1 activation in a whole-cell lysate of heat-shocked (HS) BHK cells by RNase A. The lysate (10 µg total protein) was treated with 500 ng RNase A for 1 h at 37 °C with or without excess tRNA. The amount of HSF1 in each lane was monitored by immunoblotting (lower panel). **b**, HSR1 from the eEF1A fraction. Silver-stained denaturing 4% PAGE of the RNA isolated from pooled eEF1A-containing fractions (left panel). Where indicated, HSR1 samples were treated for 30 min at 25 °C with DNase I (10 U) or RNase A (100 ng) before loading onto the gel (right panel). **c**, Reconstitution of the HSF1-activating complex. Schematic on the top shows HSR1 cDNA flanked by T7 and T3 promoters. Lower panel shows EMSA of recombinant HSF1 incubated with pure eEF1A and HSR1, which was isolated either from HeLa- or BHK-activating fractions (lanes 7, 8) or synthesized by *in vitro* transcription using T7 or T3 polymerase (lanes 9–11). Quantification of HSF1 activation is presented as the fold increase relative to a background control (lane 1). **d**, Electrophoretic mobility of *in vitro* synthesized and endogenous HSR1. HSR1 isolated from BHK cells (lane 1) or synthesized by *in vitro* transcription (T3, lanes 2–4) was incubated in Mg^{2+} free buffer (–)

(lane 4) or in buffer containing 4 mM Mg^{2+} (lanes 1–3) and subjected to electrophoresis on a 1% agarose gel. The sample in lane 2 was heated at 43 °C before loading. M, RNA ladder (nt). **e**, Co-immunoprecipitation (IP) of HSR1 with an anti-eEF1A antibody. IP was performed as described in the legend to Fig. 1e. RNA was isolated from a precipitate prepared from heat shocked (HS, lane 3) or control (lane 4) cells and subjected to RT-PCR using HSR1-specific primers. Lane 1, no reverse transcription step (–RT). Lane 2, *in vitro* transcribed HSR1 (HSR1-T3) was used in place of HSR1 isolated from the immunoprecipitate. **f**, Mapping functional domains of HSR1. Fifteen overlapping 45-mer HSR1 antisense DNA oligonucleotides covering the entire length of HSR1 were screened for their ability to inhibit HSF1 activation in the reconstituted system. Recombinant HSF1 (<10 nM) was activated in the presence of purified eEF1A (0.01 mg ml^{–1}) and HSR1-T3 (0.4 µM). Where indicated, an antisense oligonucleotide was added to a final concentration of 10 µM. Changes in HSF1 activation were quantified after setting the activity of HSF1 in the absence of oligonucleotide at 100% (lane 2) and without eEF1A at 0% (lane 1). Note that some oligonucleotides potentiate HSF1 activation (for example, oligonucleotide no. 10).

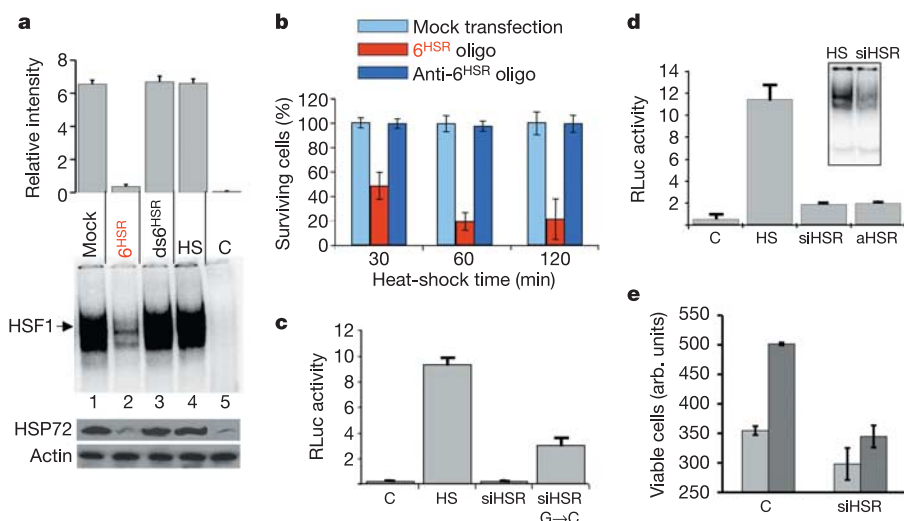


Figure 3 | Inhibition of the heat-shock response *in vivo* by targeting HSR1.

a, Effect of 6^{HSR} antisense oligonucleotide transfections on HSF1 activation and *Hsp* expression in BHK cells. Cells were transfected with 100 nM of either 6^{HSR} antisense oligonucleotide (lane 2), an equal amount of the corresponding double-stranded oligonucleotide (ds6^{HSR}, lane 3) or were mock-transfected (lane 1). Transfected cells were heat-shocked for 1 h at 43 °C, a whole-cell lysate was immediately prepared from half of the cells and 20 µg of the total protein was analysed by EMSA for HSF1 activity. The remaining half of the cells were allowed to recover at 37 °C for 16 h before a whole-cell lysate was analysed by immunoblotting using an anti-HSP72 (inducible HSP70) antibody. Control (lane 5), lysate of unstressed cells. HS (lane 4), lysate of untransfected heat-shocked cells. Upper panel shows quantification of EMSA data. Error bars represent s.d. from 3 independent experiments. **b**, Sensitizing BHK cells to heat shock with the 6^{HSR} oligonucleotide. Cells were transfected with 100 nM 6^{HSR} or anti-6^{HSR}, or were mock-transfected (–). Thermotolerance was induced by incubation for 1 h at 43 °C followed by 15 h recovery at 37 °C. Lethal heat shock (45 °C) was performed for the indicated time periods. Cell viability was determined by MTS assay 16 h after lethal heat shock. Data for 6^{HSR} and anti-6^{HSR} oligonucleotides were normalized to the efficiency of transfection. Similar results were obtained with HeLa cells. Data show mean ± s.e.m. from 6

independent experiments. **c**, Inhibition of heat shock promoter activation by siRNA against HSR1. RLuc activity was measured in lysates of BHK cells transiently transfected with a mixture of RLuc reporter plasmid, a β-gal internal control, and an siRNA construct expressing siRNA against HSR1 (siHSR) or its mutant derivative (siHSR: C → G). C, control lysate of unstressed cells. Similar results were observed with HeLa cells. Error bars represent s.d. from 4 independent experiments. **d**, Heat-shock response in stably transfected cells expressing siRNA against HSR1. RLuc activity was measured in lysates of HeLa cell lines that stably express either GFP (HS), siRNA against HSR1 (siHSR) or full-length antisense HSR1 (aHSR). Each of these cell lines was transiently transfected with an RLuc reporter plasmid and a β-gal internal control. C, lysate of unstressed cells. The HSF1 activity in the lysates of GFP- (HS) and siHSR- expressing cells is shown in the inset. Error bars represent s.d. from 4 independent experiments. **e**, Inhibition of thermotolerance by siRNA against HSR1. Cells stably expressing GFP (control, C) or siRNA against HSR1 (siHSR) were subjected to a lethal heat shock for 2 h at 46 °C with (dark grey) or without (light grey) pre-conditioning at 43 °C for 1 h followed by 16 h recovery at 37 °C. Cell viability was determined by the MTS-based assay. Data show mean ± s.e.m. from 3 independent experiments.

chaperone-based repression of HSF1 and the activation mechanism described here are likely to coexist (Supplementary Fig. 4). Thus, HSR1–eEF1A complexes could capture HSF1 released from the HSP90 complex and assist its assembly into trimers and/or increase the stability of HSF1 trimers.

Our data assign a new function to eEF1A as a co-activator of HSF1 and suggest that this ubiquitous and extremely conserved protein has a multifaceted role in the response to heat shock. Heat shock leads to two major physiological perturbations in the cell—translational shutdown and cytoskeleton collapse. These events might cause the release of eEF1A, which then becomes available for interaction with HSR1 and HSF1 to initiate the heat-shock response (Supplementary Fig. 4). Thus, eEF1A may provide a functional link between aberrations in general cellular processes and the stress response.

Recently, non-coding RNAs have been implicated in a wide spectrum of regulatory functions^{27,28}. Here we describe a previously uncharacterised non-coding RNA that is essential for heat-shock response activation and that could potentially serve as a thermosensor. Indeed, the concept of an RNA thermosensor, albeit with a different mode of action, has already been described in bacteria²⁹. HSR1 may represent an attractive pharmacological target in various pathological conditions associated with HSF1 activation, including inflammation, ischaemia/reperfusion and cancer³⁰.

METHODS

Cell culture, heat-shock treatment and lysate preparation. HeLa and BHK-21 cells were grown at 37 °C in a 5% CO₂ atmosphere in DMEM medium

containing 10% fetal bovine serum (FBS) and a DMEM/F-12 (1:1) mixture containing 10% newborn calf serum (NBCS), respectively. Growth media contained 2 mM glutamine and an antibiotic/antimycotic cocktail (penicillin/streptomycin/fungizone) (Invitrogen). Heat shock of cultured cells was performed by submerging tightly closed, parafilm-sealed screw-cap flasks in a water bath adjusted to 43 °C or 45 °C as indicated. Control cells were maintained in tightly closed flasks in an incubator at 37 °C. Whole-cell lysate was prepared by three cycles of freezing in liquid nitrogen/thawing at ambient temperature in HEDG buffer (20 mM HEPES-NaOH, pH 7.8, 0.5 mM EDTA, 0.5 mM dithiothreitol (DTT), 10% glycerol and protease inhibitor cocktail (Complete, Roche)) containing 0.42 M NaCl at a ratio of 100 µl per confluent 75-cm² flask followed by centrifugation at 25,000g for 15 min. The lysate was divided into small aliquots, flash-frozen and stored at –70 °C. Total protein concentration was ~3–5 mg ml^{–1}.

Protein expression and purification. The pGex-2T plasmid carrying mouse *Hsf1* cDNA fused to amino-terminal GST was a gift from K. Sarge. *E. coli* BL-21 cells were transformed with pGex-2THSF1, grown until they reached an absorbance of ~0.6 at 600 nm (*A*₆₀₀) and induced with 0.1 mM IPTG for 5 h at 28 °C. For details of the purification procedure and HSF1 immobilization on Sepharose, see Supplementary Methods. eEF1A was purified from rat liver essentially as described in ref. 21, and is described in detail in the Supplementary Methods.

HSF1-Sepharose pull-down experiments and isolation of HSR1. For the isolation of eEF1A fraction, whole-cell lysate was prepared as described above and diluted with HEDG buffer to give a final concentration of 0.21 M NaCl. A bed volume of 10 µl HSF1-Sepharose beads per 75-cm² flask of confluent cells was added and mixed at 4 °C for 4 h. Beads were collected by centrifugation at 500g for 5 min and washed three times with ten bed volumes of HEDG containing 0.21 M NaCl. Elution was performed by incubating the beads with

one bed volume of HEDG containing 0.21 M NaCl at 43 °C for 30 min with occasional mixing. Three successive elutions were performed and the fractions were analysed by SDS–PAGE. To isolate HSR1, eEF1A-containing fractions were pooled, supplemented with 12.5 mM EDTA and 0.25% SDS and treated with 6,000 units of proteinase K (Fermentas) for 30 min at 50 °C. RNA was then extracted twice with phenol-chloroform and precipitated with isopropanol.

EMSA and chemical crosslinking. EMSA was performed as described in ref. 2 with minor modifications. Binding reactions contained 10–20 µg protein (from whole-cell lysate) or 10 nM purified recombinant HSF1, 20 mM HEPES–NaOH (pH 7.9), 100 mM NaCl, 1.5 mM MgCl₂, 0.5 mM EDTA, 2 mM DTT, 10% (v/v) glycerol, 1 mg ml^{−1} BSA, 2.5 µg poly(dI–dC), and 1.25 ng of ³²P-end-labelled HSE oligonucleotide (5′-GCCTCGAATGTTTCGCGAAGTTTCG-3′) in a final volume of 40 µl. Reactions were incubated for 20 min at room temperature (22–25 °C), mixed with 4 µl of 0.02% bromophenol blue in 50% glycerol, loaded onto a 4% polyacrylamide gel, and run in 25 mM Tris, 190 mM glycine, 1 mM EDTA (pH 8.3) at 50 mA. The gel was dried and analysed using a phosphor-imager. Recombinant HSF1 (10 nM) was incubated with eEF1A and HSR1 at the indicated concentrations. The crosslinking reagent EGS (Pierce) was added to 0.2–0.5 mM in the same buffer used for EMSA for 30 min at room temperature. Reactions were quenched with 75 mM glycine and processed for immunoblotting.

Transfections and cell viability assay. The oligonucleotides (IDT) used for transfection, 6^{HSR} (5′-AGTCCTCACACCACTCCAACGCTGCTTATGGCGGCGGATTCAAC-3′) and anti-6^{HSR} (5′-GTTGAATCCGCCGCCATAAGCAGACGTTGGAGTGGTGTGAGGACT-3′), were phosphothioate-modified at the 3′ and 5′ ends to increase their stability. In the experiment shown in Fig. 3a, 6^{HSR} oligonucleotide or 6^{HSR}/anti-6^{HSR} ds-oligonucleotide (0.56 µM final concentration) were used to transfect BHK cells using Superfect reagent (Qiagen) according to the manufacturer's instructions. The next day, cells were heat-shocked for 1 h at 43 °C and a whole-cell lysate was prepared to determine HSF1 activity. In some experiments, cells were allowed to recover at 37 °C overnight after heat shock to allow synthesis of HSPs. Next, a whole-cell lysate was prepared and analysed by immunoblotting using anti-HSP72 antibodies (Stressgen) that specifically recognize only the inducible form of HSP70. In the experiment depicted in Fig. 3b, cells were transfected using Oligofectamine reagent (Invitrogen) at a final oligonucleotide concentration of 100 nM, according to the manufacturer's instructions. Cell viability was determined using the MTS reagent-based CellTiter 96 Aqueous non-radioactive cell proliferation assay (Promega, Supplementary Methods). For siRNA experiments, cells were transfected in 12-well plates with a mixture of 100 ng pMLucHSP70, 50 ng pSV40βGal (Supplementary Methods) and 250 ng siRNA construct using Effectene reagent (Qiagen), according to the manufacturer's instructions. Cells were incubated with the transfection complexes for 24 h, heat-shocked for 2 h at 43 °C and allowed to recover at 37 °C for 16 h. See Supplementary Methods for details of the generation of stably transfected cell lines. Transfection with RLuc and β-gal reporter plasmids, heat-shock treatment, and reporter gene activity assays were performed as described above for transient transfection.

Received 31 August; accepted 15 December 2005.

- Sarge, K. D., Zimarino, V., Holm, K., Wu, C. & Morimoto, R. I. Cloning and characterization of two mouse heat shock factors with distinct inducible and constitutive DNA-binding ability. *Genes Dev.* **5**, 1902–1911 (1991).
- Sarge, K. D., Murphy, S. P. & Morimoto, R. I. Activation of heat shock gene transcription by heat shock factor 1 involves oligomerization, acquisition of DNA-binding activity, and nuclear localization and can occur in the absence of stress. *Mol. Cell. Biol.* **13**, 1392–1407 (1993).
- Westwood, J. T. & Wu, C. Activation of *Drosophila* heat shock factor: conformational change associated with a monomer-to-trimer transition. *Mol. Cell. Biol.* **13**, 3481–3486 (1993).
- Zou, J., Guo, Y., Guettouche, T., Smith, D. F. & Voellmy, R. Repression of heat shock transcription factor HSF1 activation by HSP90 (HSP90 complex) that forms a stress-sensitive complex with HSF1. *Cell* **94**, 471–480 (1998).
- Voellmy, R. Transcriptional regulation of the metazoan stress protein response. *Prog. Nucleic Acid Res. Mol. Biol.* **78**, 143–185 (2004).
- Shi, Y., Mosser, D. D. & Morimoto, R. I. Molecular chaperones as HSF1-specific transcriptional repressors. *Genes Dev.* **12**, 654–666 (1998).
- Abravaya, K., Myers, M. P., Murphy, S. P. & Morimoto, R. I. The human heat shock protein hsp70 interacts with HSF, the transcription factor that regulates heat shock gene expression. *Genes Dev.* **6**, 1153–1164 (1992).
- Marchler, G. & Wu, C. Modulation of *Drosophila* heat shock transcription factor activity by the molecular chaperone DROJ1. *EMBO J.* **20**, 499–509 (2001).
- Zimarino, V., Tsai, C. & Wu, C. Complex modes of heat shock factor activation. *Mol. Cell. Biol.* **10**, 752–759 (1990).
- Shopland, L. S., Hirayoshi, K., Fernandes, M. & Lis, J. T. HSF access to heat shock elements *in vivo* depends critically on promoter architecture defined by GAGA factor, TFIID, and RNA polymerase II binding sites. *Genes Dev.* **9**, 2756–2769 (1995).
- Shopland, L. S. & Lis, J. T. HSF recruitment and loss at most *Drosophila* heat shock loci is coordinated and depends on proximal promoter sequences. *Chromosoma* **105**, 158–171 (1996).
- Zuo, J., Rungger, D. & Voellmy, R. Multiple layers of regulation of human heat shock transcription factor 1. *Mol. Cell. Biol.* **15**, 4319–4330 (1995).
- Ali, A., Bharadwaj, S., O'Carroll, R. & Ovsenek, N. HSP90 interacts with and regulates the activity of heat shock factor 1 in *Xenopus* oocytes. *Mol. Cell. Biol.* **18**, 4949–4960 (1998).
- Bharadwaj, S., Ali, A. & Ovsenek, N. Multiple components of the HSP90 chaperone complex function in regulation of heat shock factor 1 *in vivo*. *Mol. Cell. Biol.* **19**, 8033–8041 (1999).
- Clos, J., Rabindran, S., Wisniewski, J. & Wu, C. Induction temperature of human heat shock factor is reprogrammed in a *Drosophila* cell environment. *Nature* **364**, 252–255 (1993).
- Batulan, Z. *et al.* High threshold for induction of the stress response in motor neurons is associated with failure to activate HSF1. *J. Neurosci.* **23**, 5789–5798 (2003).
- Gothard, L. Q., Ruffner, M. E., Woodward, J. G., Park-Sarge, O. K. & Sarge, K. D. Lowered temperature set point for activation of the cellular stress response in T-lymphocytes. *J. Biol. Chem.* **278**, 9322–9326 (2003).
- Jolly, C., Usson, Y. & Morimoto, R. I. Rapid and reversible relocation of heat shock factor 1 within seconds to nuclear stress granules. *Proc. Natl. Acad. Sci. USA* **96**, 6769–6774 (1999).
- Panniers, R. Translational control during heat shock. *Biochimie* **76**, 737–747 (1994).
- Welch, W. J. & Suhan, J. P. Morphological study of the mammalian stress response: characterization of changes in cytoplasmic organelles, cytoskeleton, and nucleoli, and appearance of intranuclear actin filaments in rat fibroblasts after heat-shock treatment. *J. Cell Biol.* **101**, 1198–1211 (1985).
- Kristensen, P., Lund, A., Clark, B. F., Cavallius, J. & Merrick, W. C. Purification and characterisation of a tissue specific elongation factor 1 alpha (EF-1α2) from rabbit muscle. *Biochem. Biophys. Res. Commun.* **245**, 810–814 (1998).
- Baler, R., Zou, J. & Voellmy, R. Evidence for a role of Hsp70 in the regulation of the heat shock response in mammalian cells. *Cell Stress Chaperones* **1**, 33–39 (1996).
- Farkas, T., Kutsikova, Y. A. & Zimarino, V. Intramolecular repression of mouse heat shock factor 1. *Mol. Cell. Biol.* **18**, 906–918 (1998).
- Zhong, M., Orosz, A. & Wu, C. Direct sensing of heat and oxidation by *Drosophila* heat shock transcription factor. *Mol. Cell* **2**, 101–108 (1998).
- Zhong, M., Kim, S. J. & Wu, C. Sensitivity of *Drosophila* heat shock transcription factor to low pH. *J. Biol. Chem.* **274**, 3135–3140 (1999).
- Clos, J. *et al.* Molecular cloning and expression of a hexameric *Drosophila* heat shock factor subject to negative regulation. *Cell* **63**, 1085–1097 (1990).
- Szymanski, M. & Barciszewski, J. Regulation by RNA. *Int. Rev. Cytol.* **231**, 197–258 (2003).
- Nudler, E. & Mironov, A. S. The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.* **29**, 11–17 (2004).
- Storz, G. An RNA thermometer. *Genes Dev.* **13**, 633–636 (1999).
- Westerheide, S. D. & Morimoto, R. I. Heat shock response modulators as therapeutic tools for diseases of protein conformation. *J. Biol. Chem.* **280**, 33097–33100 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Schick and E. Avetisova for technical assistance and N. J. Cowan for critical reading of the manuscript. This work was supported by grants from the NIH and the Edward Mallinckrodt Jr Foundation (E.N.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.N. (evgeny.nudler@med.nyu.edu) or D.G. (dgershon@redoxpharm.com).

Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation

Meenakshi K. Doma¹ & Roy Parker¹

A fundamental aspect of the biogenesis and function of eukaryotic messenger RNA is the quality control systems that recognize and degrade non-functional mRNAs. Eukaryotic mRNAs where translation termination occurs too soon (nonsense-mediated decay)¹ or fails to occur (non-stop decay)² are rapidly degraded. We show that yeast mRNAs with stalls in translation elongation are recognized and targeted for endonucleolytic cleavage, referred to as 'no-go decay'. The cleavage triggered by no-go decay is dependent on translation and involves Dom34p and Hbs1p. Dom34p and Hbs1p are similar to the translation termination factors eRF1 and eRF3 (refs 3, 4), indicating that these proteins might function in recognizing the stalled ribosome and triggering endonucleolytic cleavage. No-go decay provides a mechanism for clearing the cell of stalled translation elongation complexes, which could occur as a result of damaged mRNAs or ribosomes, or as a mechanism of post-transcriptional control.

To determine the effect of a delay in translation elongation on mRNA decay, we inserted a stable stem-loop, previously shown to inhibit translation elongation⁵, into the coding region of a yeast PGK1 reporter construct⁶ to give the PGK1-SL reporter. We observed that the PGK1-SL mRNA decays faster than the control PGK1 mRNA in wild-type yeast cells (Supplementary Fig. 1), suggesting that a ribosomal stall accelerated degradation of the mRNA.

In yeast, the two major pathways of mRNA turnover initiate with deadenylation of the 3'-polyadenosine (poly(A)) tail⁷. Deadenylation is followed either by decapping and 5' → 3' degradation, or 3' → 5' degradation⁷, of the mRNA. The PGK1-SL mRNA showed accelerated decay in both a *dcp2Δ* strain, which lacks decapping activity⁸, and a *ski7Δ* strain, which is defective in the 3' → 5' degradation pathway² (Supplementary Fig. 1). Moreover, the PGK1-SL mRNA showed the same instability in an *upf1Δ* strain, which is defective for the nonsense-mediated decay (NMD) pathway¹ (Supplementary Fig. 1). Thus, the PGK1-SL mRNA shows accelerated degradation in a manner independent of known degradation pathways, suggesting that it is targeted for degradation by a novel mechanism, possibly involving cleavage by an endonuclease.

If the PGK1-SL mRNA is subject to cleavage by an endonuclease, two mRNA fragments should be produced, a 5' fragment expected to be a substrate for the 3' → 5' degradation complex termed the exosome, and a 3' fragment that would be uncapped and predicted to be a substrate for Xrn1p, the major 5' → 3' exonuclease. Given this, we examined the PGK1-SL mRNA in strains defective in either exosome-mediated decay or Xrn1p function, using probes specific to the 5' and 3' portions of the mRNA.

Strains defective in cytoplasmic 3' → 5' degradation accumulated a fragment of the PGK1-SL mRNA detected with a probe 5' of the stem-loop (Fig. 1a). This 5' fragment accumulated in strains defective in 3' → 5' decay for any of three reasons: first, the loss of any component of the heterotrimeric Ski complex of Ski2, Ski3 and Ski8 proteins (*ski2Δ*, *ski3Δ* or *ski8Δ*), which is required for

cytoplasmic exosome function⁹; second, the loss of Ski7p (*ski7Δ*), which couples the Ski complex to the exosome; and last, a point mutation in an exosome subunit (*ski4-1*)¹⁰ that blocks the interaction of the exosome with Ski7p (data not shown). This mRNA fragment was about 1,200 nucleotides long, which is the distance from the cap to the site of the stem-loop. These observations suggest that the PGK1-SL mRNA is subject to an endonucleolytic cleavage,

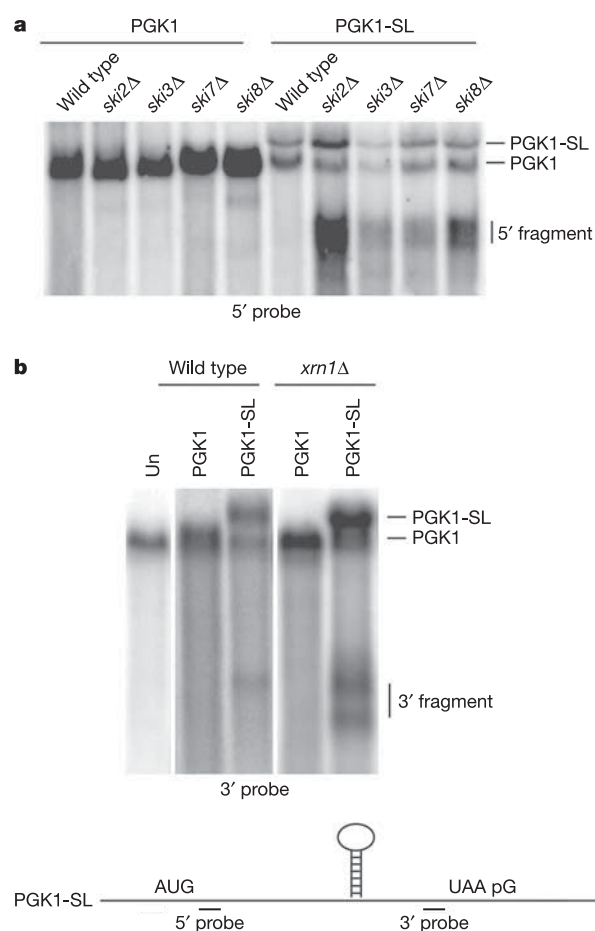


Figure 1 | Stalling of elongating ribosomes leads to endonucleolytic cleavage of the PGK1-SL reporter transcript. Northern analysis of steady-state PGK1 and PGK1-SL reporter mRNA in wild-type strains and strains mutant either in 3' → 5' exosome activity (a) that include *ski2Δ*, *ski3Δ*, *ski7Δ* and *ski8Δ*, or a mutant in 5' → 3' exonucleolytic activity, *xrn1Δ* (b). The untransformed strain (Un) is shown in b. Northern analysis was performed with probes specific for mRNA regions 5' (a) or 3' (b) of the stall site, as illustrated at the bottom.

¹Howard Hughes Medical Institute, Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85721, USA.

with the 5' mRNA fragment being degraded by the cytoplasmic exosome.

Additional evidence that the PGK1-SL mRNA is subject to endonucleolytic cleavage is that a probe specific for the 3' end of the PGK1-SL reporter detected 3' decay intermediates in the *xrn1Δ* strain (Fig. 1b). The larger fragment is about 600 nucleotides long and corresponds to the distance from the stem-loop to the 3' end of the mRNA. The smaller fragment is an mRNA fragment that has subsequently been trimmed at the 3' end to a poly(G) tract present in the PGK1 3' untranslated region, which blocks further 3' → 5' degradation¹⁰ (data not shown, and see below). The roughly 600-nucleotide 3' mRNA fragment is present at lower levels in wild-type cells and mutant *ski* strains (Fig. 1a, and Supplementary Fig. 2a), which indicates that the inserted stem-loop structure can act as a block to Xrn1p (see below). We cannot formally determine whether the mRNA fragment accumulating in the wild-type strain arises by decapping and 5' → 3' degradation by Xrn1p, or by endonucleolytic cleavage. However, the mRNA fragment detected in the *xrn1Δ* strain must be produced by endonucleolytic cleavage, arguing that at least some of the 3' decay product seen in the wild-type strain is produced by endonucleolytic cleavage. A probe specific to the 5' end of the PGK1-SL mRNA did not identify any fragment that specifically accumulated in the *xrn1Δ* strain (Supplementary Fig. 2b).

The accumulation of a 3' mRNA fragment in *xrn1Δ* strains, and a 5' mRNA fragment in *skiΔ* mutants, indicated that the PGK1-SL mRNA was endonucleolytically cleaved followed by exonuclease digestion of the mRNA fragments. We refer to this surveillance mechanism as 'no-go decay' (NGD). Insertion of the stem-loop into the coding region of the MFA2 mRNA (MFA2-SL) also identified 5' and 3' decay intermediates in the *ski7Δ* and *xrn1Δ* mutants, respectively (Supplementary Fig. 3). Mapping of the 5' and 3' fragments for both the MFA2-SL and PGK1-SL transcripts indicates that the cleavage occurs in the vicinity of the stem-loop (Supplementary Fig. 3), although there is heterogeneity indicating that the cleavage occurs at multiple sites near the ribosomal stall site and/or that there is trimming by exonucleases from the original cleavage site(s). In addition, transcriptional pulse-chase experiments and transcriptional induction experiments indicate that the mRNA fragments are produced after the full-length mRNA has been synthesized, which is consistent with a precursor-product relationship

between the full-length mRNA and the mRNA decay product (Supplementary Fig. 4, and data not shown).

Two observations indicate that deadenylation is not required to occur before endonucleolytic cleavage in NGD (Supplementary Fig. 5): first, in a transcriptional induction experiment, we observed that the 3' mRNA fragment of the MFA2-SL mRNA was polyadenylated when first produced; and second, loss of the major mRNA deadenylase, Ccr4p, does not prevent NGD.

Other stalls in translation, including a pseudoknot, rare codons and premature translation termination codons, also induce cleavage events at low temperatures, although at greatly reduced levels (Fig. 2a). This observation supports the argument that the trigger for NGD might be a delay in translation elongation.

If stalls in translation elongation trigger NGD, ribosomes should be required to reach the stem-loop to trigger the cleavage event. To test this prediction, we prevented ribosomes from reaching the stall site by blocking ribosome scanning to the AUG by inserting a stem-loop in the 5' untranslated region of the PGK1-SL mRNA¹¹ (SL-PGK1-SL). This block to translation initiation prevented the accumulation of the 5' and 3' decay intermediates in the *ski7Δ* and *xrn1Δ* strains (Fig. 3; SL-PGK1-SL). In addition, the accumulation of the 5' and 3' fragments in *ski7Δ* and *xrn1Δ* strains was abolished when a premature stop codon was inserted in the PGK1-SL mRNA at codon 22 (PGK1-PTC-SL), such that ribosomes terminated translation before the ribosomal stall site¹² (Fig. 3; PGK1-PTC-SL). Because the PGK1-PTC-SL mRNA is a substrate for NMD, its levels are quite low. To verify that these low levels of mRNA did not obscure

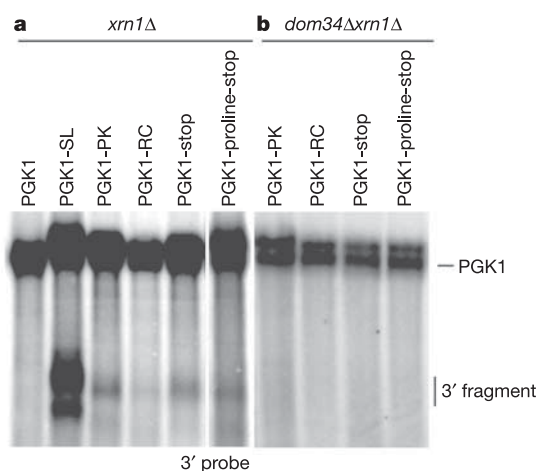


Figure 2 | NGD is initiated by endonucleolytic cleavage of mRNA with various ribosomal stalls. Steady-state cultures of *xrn1Δ* (a) and *dom34Δxrn1Δ* (b) strains with PGK1, PGK1-SL, PGK1-PK (pseudoknot), PGK1-RC (rare codons), PGK1-stop (stop codon) and PGK1-proline-stop (proline-stop codon) reporters were grown at low temperature (16°C). Northern analysis of steady-state reporter mRNAs was performed with a probe specific for mRNA regions 3' of the stall sites.

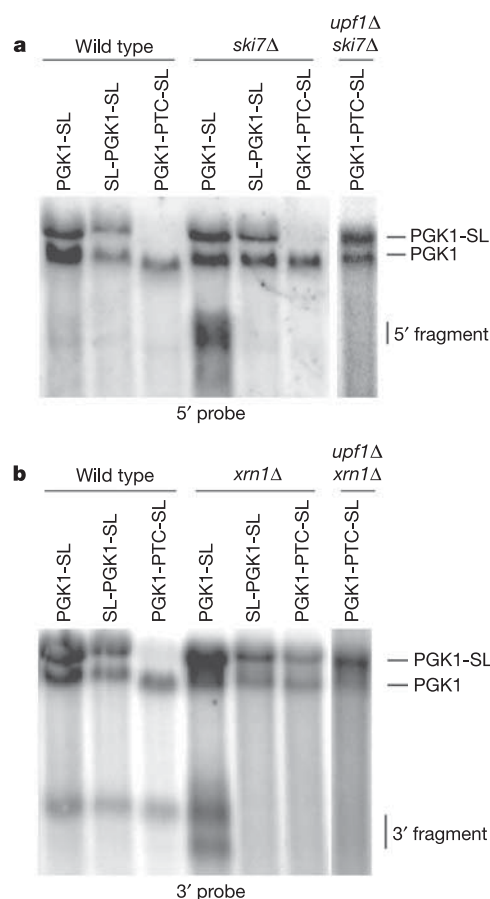


Figure 3 | NGD is dependent on translation by ribosomes. Northern analysis of steady-state PGK1-SL, SL-PGK1-SL (pRP1252) and PGK1-PTC-SL (pRP1253) reporter transcripts in wild-type, *ski7Δ*, *upf1Δ ski7Δ* (a) and *xrn1Δ*, *upf1Δ xrn1Δ* (b) mutant strains. Northern analysis was performed with probes specific for mRNA regions 5' (a) and 3' (b) of the stall site.

the accumulation of the NGD products, we also analysed this mRNA in *upf1Δski7Δ* and *upf1Δxrn1Δ* strains, in which NMD is inactivated and mRNA levels are higher. Again, no 5' or 3' cleavage products accumulated (Fig. 3). Thus, the endonucleolytic cleavage we detect requires ribosomes to translate to the stem-loop, suggesting that ribosome stalling leads to cleavage.

Because NGD works on stalled ribosomes, we proposed that Hbs1p and Dom34p, two conserved and interacting proteins with similarity to translation termination factors, might function in NGD. Hbs1p is likely to interact with the ribosome because it is a member of the family of GTPases consisting of eEF1A, which delivers the transfer RNA to the A site of the ribosome¹³, eRF3, which functions in translation termination¹⁴, and Ski7p, which is thought to interact with an empty A site when the ribosome reaches the 3' end of a mRNA². Moreover, Dom34p binds Hbs1p (ref. 15) and is related to eRF1, which has a three-dimensional structure similar to a tRNA and functions with eRF3 during translation termination¹⁶.

To test whether Hbs1p and/or Dom34p function in NGD, we determined whether the loss of Hbs1p or Dom34p affected the accumulation of NGD decay products from the PGK1-SL mRNA easily seen in *ski7Δ* and *xrn1Δ* strains. In both *dom34Δski7Δ* and *dom34Δxrn1Δ* strains the mRNA fragments generated by NGD were no longer observed (Fig. 4), indicating that Dom34p is required

for NGD. Similar results were seen with other translational stalls (Fig. 2b).

Hbs1p also has a function in NGD, because accumulation of 5' fragments from the PGK1-SL mRNA observed in *ski7Δ* mutants is substantially reduced in an *hbs1Δski7Δ* or *hbs1Δski2Δ* strain (Fig. 4a). Surprisingly, and in contrast to the *dom34Δxrn1Δ* strain, the *hbs1Δxrn1Δ* strain still showed the accumulation of the 3' fragment. This observation could be explained by either of two related possibilities. First, in an *xrn1Δ* strain, Ski7p might be able to substitute for the role of Hbs1p, because Ski7p and Hbs1p have recently diverged¹⁷. Second, Hbs1p might be required for NGD only when the cytoplasmic exosome is compromised, as seen in the *hbs1Δski7Δ* and *hbs1Δski2Δ* strains, perhaps because of a loss of Ski7p function in those mutant strains.

The above observations suggest that Dom34p recognizes stalled ribosomes, possibly in conjunction with Hbs1p, leading to endonucleolytic cleavage of the mRNA. It remains to be determined whether Dom34p is the endonuclease, activates a nuclease activity within the ribosome or recruits an endonuclease. Moreover, the identification of Dom34p and Hbs1p as mediators of NGD adds to the family of related proteins thought to modulate different events at the A site of the ribosome.

NGD provides a mechanism to release stalled, or non-functional, ribosomes and degrade damaged mRNA. An interesting possibility is that NGD has been adapted for the regulation of specific mRNAs, and previous examples of regulated ribosome-dependent cleavage of mRNAs in eukaryotes^{18,19} may be due to an inhibition of translation elongation, which then activates NGD. Consistent with this model is the observation that the sequences triggering endonucleolytic cleavage of the *CGS1* mRNA in *Arabidopsis* have been recently shown to mediate ribosome stalling²⁰. Finally, elongation pauses can also trigger endonucleolytic cleavage in prokaryotes^{21–24}, indicating that mRNA surveillance based on translation elongation rates is broadly conserved.

METHODS

The stem-loop structure (SL) was constructed by using a synthetic oligonucleotide with the following sequence and is similar to one used in an earlier study⁵: 5'-GAT ATC CCG TGG AGG GGC GCG TGG TGG CGG CTG CAG CCG CCA CCA CGC GCC CCT CCA CGG GAT ATC-3'. This was introduced into the pJD375 construct²⁵ between *SalI* and *BamHI* sites in the multiple cloning regions to give pRP1250. To generate the reporters used in this study, 275-base-pair fragments containing the SL was amplified by polymerase chain reaction with the use of primers containing *XbaI* ends and ligated into the *XbaI* site of wild-type PGK1 (pRP469) to give the PGK1-SL construct (pRP1251). Insertion of other translational stalls in the PGK1 mRNA, including the pseudoknot²⁶, rare codons²⁷, stop codon and proline-stop codon²⁸ was done in the same manner, generating PGK1-PK(pRP1285), PGK1-RC(pRP1286), PGK1-stop(pRP1287) and PGK1-proline-stop(pRP1288) reporters. Primer sequences are available from the authors on request. The reporters used for the translational dependence of NGD were constructed by introduction of the *XbaI*-generated fragment containing the SL into pRP543 (ref. 11) and pRP609 (ref. 12) to give pRP1252 and pRP1253, respectively. All PGK1 constructs are under the control of the *GalI* upstream activation sequence (UAS). Yeast strains are detailed in Supplementary Information.

All RNA analyses were performed as described previously¹². All experiments were performed at 30 °C in galactose-containing medium with mid-log cultures except for those in the experiments in Fig. 2, which were grown at 16 °C. For the PGK1 reporter mRNAs, the oligonucleotide oRP132 was used for detection of the 5' fragment and the oligonucleotide oRP70 for detection of the 3' fragment. Loading corrections for quantification were determined by hybridization with oRP100, an oligonucleotide directed to *SCR1* RNA, a stable RNA polymerase III transcript.

Received 5 August; accepted 13 December 2005.

- Maquat, L. E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature Rev. Mol. Cell Biol.* 5, 89–99 (2004).
- van Hoof, A., Frischmeyer, P. A., Dietz, H. C. & Parker, R. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science* 295, 2262–2264 (2002).

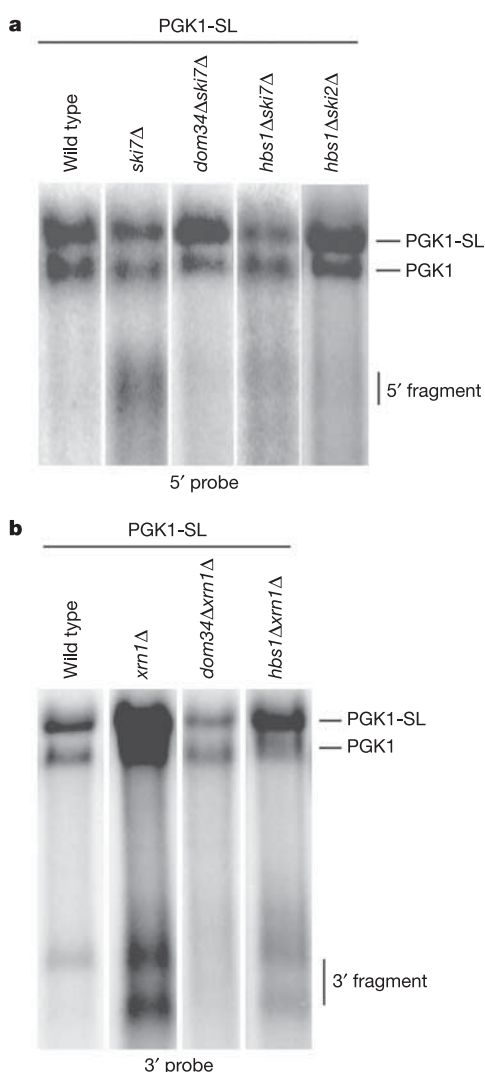


Figure 4 | Dom34p and Hbs1p affect NGD. Northern analysis of steady-state PGK1-SL reporter mRNA in wild-type and mutant strains with probes specific for mRNA regions 5' (a) and 3' (b) of the stall site.

3. Davis, L. & Engebrecht, J. Yeast dom34 mutants are defective in multiple developmental pathways and exhibit decreased levels of polyribosomes. *Genetics* **149**, 45–56 (1998).
4. Inagaki, Y., Blouin, C., Susko, E. & Roger, A. J. Assessing functional divergence in EF-1 α and its paralogs in eukaryotes and archaeobacteria. *Nucleic Acids Res.* **31**, 4227–4237 (2003).
5. Hosoda, N. *et al.* Translation termination factor eRF3 mediates mRNA decay through the regulation of deadenylation. *J. Biol. Chem.* **278**, 38287–38291 (2003).
6. Decker, C. J. & Parker, R. A turnover pathway for both stable and unstable mRNAs in yeast: evidence for a requirement for deadenylation. *Genes Dev.* **7**, 1632–1643 (1993).
7. Collier, J. & Parker, R. Eukaryotic mRNA decapping. *Annu. Rev. Biochem.* **73**, 861–890 (2004).
8. Duncley, T. & Parker, R. The DCP2 protein is required for mRNA decapping in *Saccharomyces cerevisiae* and contains a functional MutT motif. *EMBO J.* **18**, 5411–5422 (1999).
9. van Hoof, A., Staples, R. R., Baker, R. E. & Parker, R. Function of the ski4p (Csl4p) and Ski7p proteins in 3'-to-5' degradation of mRNA. *Mol. Cell. Biol.* **20**, 8230–8243 (2000).
10. Anderson, J. S. & Parker, R. The 3' to 5' degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3' to 5' exonucleases of the exosome complex. *EMBO J.* **17**, 1497–1506 (1998).
11. Muhlrud, D., Decker, C. J. & Parker, R. Turnover mechanisms of the stable yeast PGK1 mRNA. *Mol. Cell. Biol.* **15**, 2145–2156 (1995).
12. Muhlrud, D. & Parker, R. Premature translational termination triggers mRNA decapping. *Nature* **370**, 578–581 (1994).
13. Inge-Vechtomov, S., Zhouravleva, G. & Philippe, M. Eukaryotic release factors (eRFs) history. *Biol. Cell.* **95**, 195–209 (2003).
14. Nelson, R. J., Ziegelhoffer, T., Nicolet, C., Werner-Washburne, M. & Craig, E. A. The translation machinery and 70 kd heat shock protein cooperate in protein synthesis. *Cell* **71**, 97–105 (1992).
15. Carr-Schmid, A., Pfund, C., Craig, E. A. & Kinzy, T. G. Novel G-protein complex whose requirement is linked to the translational status of the cell. *Mol. Cell. Biol.* **22**, 2564–2574 (2002).
16. Kong, C. *et al.* Crystal structure and functional analysis of the eukaryotic class II release factor eRF3 from *S. pombe*. *Mol. Cell* **14**, 233–245 (2004).
17. van Hoof, A. Conserved functions of yeast genes support the Duplication, Degeneration and Complementation model for gene duplication. *Genetics* **171**, 1455–1461 (2005).
18. Chiba, Y. *et al.* Evidence for autoregulation of cystathionine gamma-synthase mRNA stability in *Arabidopsis*. *Science* **286**, 1371–1374 (1999).
19. Gattfield, D. & Izaurralde, E. Nonsense-mediated messenger RNA decay is initiated by endonucleolytic cleavage in *Drosophila*. *Nature* **429**, 575–578 (2004).
20. Onouchi, H. *et al.* Nascent peptide-mediated translation elongation arrest coupled with mRNA degradation in the CGS1 gene of *Arabidopsis*. *Genes Dev.* **19**, 1799–1810 (2005).
21. Sunohara, T., Jojima, K., Tagami, H., Inada, T. & Aiba, H. Ribosome stalling during translation elongation induces cleavage of mRNA being translated in *Escherichia coli*. *J. Biol. Chem.* **279**, 15368–15375 (2004).
22. Sunohara, T., Jojima, K., Yamamoto, Y., Inada, T. & Aiba, H. Nascent-peptide-mediated ribosome stalling at a stop codon induces mRNA cleavage resulting in nonstop mRNA that is recognized by tmRNA. *RNA* **10**, 378–386 (2004).
23. Hayes, C. S. & Sauer, R. T. Cleavage of the A site mRNA codon during ribosome pausing provides a mechanism for translational quality control. *Mol. Cell* **12**, 903–911 (2003).
24. Withey, J. H. & Friedman, D. I. A salvage pathway for protein structures: tmRNA and trans-translation. *Annu. Rev. Microbiol.* **57**, 101–123 (2003).
25. Harger, J. W. & Dinman, J. D. An *in vivo* dual-luciferase assay system for studying translational recoding in the yeast *Saccharomyces cerevisiae*. *RNA* **9**, 1019–1024 (2003).
26. Kozak, M. Constraints on reinitiation of translation in mammals. *Nucleic Acids Res.* **29**, 5226–5232 (2001).
27. Sorensen, M. A., Kurland, C. G. & Pedersen, S. Codon usage determines translation rate in *Escherichia coli*. *J. Mol. Biol.* **207**, 365–377 (1989).
28. Hayes, C. S., Bose, B. & Sauer, R. T. Proline residues at the C terminus of nascent chains induce SsrA tagging during translation termination. *J. Biol. Chem.* **277**, 33825–33832 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Farabaugh for the pJD375 plasmid; A. van Hoof for strains; and members of the Parker laboratory, especially K. Baker and C. Decker, for discussions. This study was supported by funds from the Howard Hughes Medical Institute and the National Institutes of Health.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.P. (rrparker@u.arizona.edu).

The Dam1 kinetochore ring complex moves processively on depolymerizing microtubule ends

Stefan Westermann¹, Hong-Wei Wang², Agustin Avila-Sakar², David G. Drubin¹, Eva Nogales^{1,2,3} & Georjana Barnes¹

Chromosomes interact through their kinetochores with microtubule plus ends and they are segregated to the spindle poles as the kinetochore microtubules shorten during anaphase A of mitosis. The molecular natures and identities of coupling proteins that allow microtubule depolymerization to pull chromosomes to poles during anaphase have long remained elusive¹. In budding yeast, the ten-protein Dam1 complex is a critical microtubule-binding component of the kinetochore² that oligomerizes into a 50-nm ring around a microtubule *in vitro*^{3,4}. Here we show, with the use of a real-time, two-colour fluorescence microscopy assay, that the ring complex moves processively for several micrometres at the ends of depolymerizing microtubules without detaching from the lattice. Electron microscopic analysis of 'end-on views' revealed a 16-fold symmetry of the kinetochore rings. This out-of-register arrangement with respect to the 13-fold microtubule symmetry is consistent with a sliding mechanism based on an electrostatically coupled ring-microtubule interface. The Dam1 ring complex is a molecular device that can translate the force generated by microtubule depolymerization into movement along the lattice to facilitate chromosome segregation.

Although in principle minus-end-directed motor proteins could be responsible for the poleward transport of chromosomes, it has long been recognized that the force generated by microtubule depolymerization itself contributes to movements during anaphase A^{1,5}. Microtubule depolymerization is sufficient to move metazoan chromosomes⁶ or kinesin-coated microbeads *in vitro*^{7,8}. Different theoretical models have tried to explain how kinetochores exploit the force created by depolymerization^{9,10}, but the molecular nature of a device that would remain attached to a disassembling microtubule end without losing its connection has remained elusive.

Budding yeast mitosis displays anaphase A and B movements¹¹, and kinetochore microtubule dynamics are restricted to plus ends¹². Recent studies have revealed that the budding yeast kinetochore is composed of at least eight multiprotein complexes that assemble on centromeric DNA in a hierarchical manner^{13–15}. A critical microtubule binding activity of the kinetochore is provided by the ten-protein Dam1 complex, which is essential for mitotic spindle integrity and, under regulation by the yeast Aurora B kinase Ipl1p, for the establishment and maintenance of bipolar chromosome attachments^{16–18}.

Purified Dam1 complex oligomerizes into 50-nm ring structures when incubated with microtubules *in vitro*^{3,4}. The Dam1 rings form at physiologically relevant concentrations, encircle the microtubule orthogonal to its axis, promote microtubule assembly, and preferentially bind to the GTP lattice of GTP/GDP-segmented microtubules, suggesting a mechanism for their targeting to plus ends³. Two observations suggested that the ring complex might exhibit lateral

mobility on the microtubule: first, the nature of the interaction between rings and microtubules is largely electrostatic, and second, electron micrographs of the Dam1 ring complex showed that rings accumulated at the ends of depolymerizing microtubules³.

To uncover the principles of force generation and motility at the kinetochore we sought to develop a fluorescence microscopy assay that would allow observation of the Dam1 complex on dynamic microtubules in real time. The main challenge was to observe microtubules that are not completely attached to the surface of a flow cell because this might impede movement of the ring complex. This was achieved by imaging the microtubules in a flow chamber filled with a viscous solution of methylcellulose and by blocking the glass surfaces to limit non-specific adsorption¹⁹ (Fig. 1a). To mimic microtubule dynamics during anaphase we induced depolymerization either by including the kinesin13 family member XMCAK1 in the motility buffer or by diluting Dam1-decorated microtubules below the critical concentration for microtubule assembly. Indistinguishable results were obtained by both disassembly protocols. When 5 nM Alexa488-conjugated Dam1 complex was incubated with rhodamine-labelled microtubules, which were then diluted, microtubules depolymerized over the course of several minutes. As the free end of the microtubule shortened, the Alexa488 signal could clearly be seen tracking with the disassembling end (Fig. 1b and Supplementary Movie S1). In cases where the observed microtubule was completely free in solution, Dam1 signals tracked both depolymerizing ends and increased in intensity as the microtubule shortened (Supplementary Movie S2). In rare events, the Dam1 signal was partly released from the depolymerizing end, probably as a result of adsorption of the complex to the coverslip surface. In these cases, the green fluorescent Dam1 signal was 'stretched' before release, demonstrating the force generated by the depolymerizing microtubule end (Supplementary Movie S3).

In principle, tracking could occur through one of two different mechanisms: continued association of an intact ring with the microtubule lattice ('sliding'), or the rapid dissociation and reassociation of ring subunits with the retracting microtubule end ('turn-over'). To distinguish between these possibilities, we mixed green fluorescent Alexa488–Dam1 complex with red fluorescent Alexa594–Dam1 complex and incubated this mixture with unlabelled microtubules; after incubation for 30 min we visualized them in our assay system. Because rings were formed by randomly combining green and red fluorescent complexes, each spot on the microtubule showed a strict co-localization of green and red signals (Fig. 1c). In the next experiment, we added green and then red fluorescent Dam1 complex sequentially to microtubules. If the rings turned over, one would expect the subunits to re-equilibrate such that the spots on the microtubules would be green and red as in Fig. 1c. However, in this

¹Department of Molecular and Cell Biology, ²Lawrence Berkeley National Laboratory, ³Howard Hughes Medical Institute, University of California Berkeley, Berkeley, California 94720-3202, USA.

experiment most spots on the microtubule were either red or green (Fig. 1d). The rare cases in which red and green signals seemed to co-localize probably arose from adjacent rings on the microtubule too close to be resolved by fluorescence microscopy. We conclude that under our assay conditions, once bound to the microtubule, the ring subunits do not turn over and the ring tracks with the depolymerizing microtubule end as an intact entity.

We next tested whether the ring complex can couple microtubule depolymerization to the movement of a cargo. We prepared modified Dam1 rings by combining green fluorescent complexes with biotinylated complexes. These Dam1 rings readily recruited streptavidin-

coated, red fluorescent microbeads to microtubules (Supplementary Fig. 1). After the induction of depolymerization, the beads remained attached to Dam1-decorated depolymerizing microtubule ends and were translocated in the direction of depolymerization (Fig. 1e and Supplementary Movie S4). This result shows that the ring complex can act as a coupling device under load to move objects with the depolymerizing microtubule end.

Next we investigated the behaviour of multiple well-spaced Dam1 ring complexes on depolymerizing microtubule ends. As the microtubule disassembled, the distal Dam1 signal tracked with the end and merged with the next signal; both continued to move together with increased fluorescence intensity (Fig. 2a and Supplementary Movie 5). This 'collection' of Dam1 signals is further illustrated by a kymograph representation (Fig. 2b), which shows oblique Dam1 lines that merge and increase in intensity as a function of time. We quantified the Alexa488 signal of the depolymerizing end and found that it increased in steps of approximately equal magnitude. In

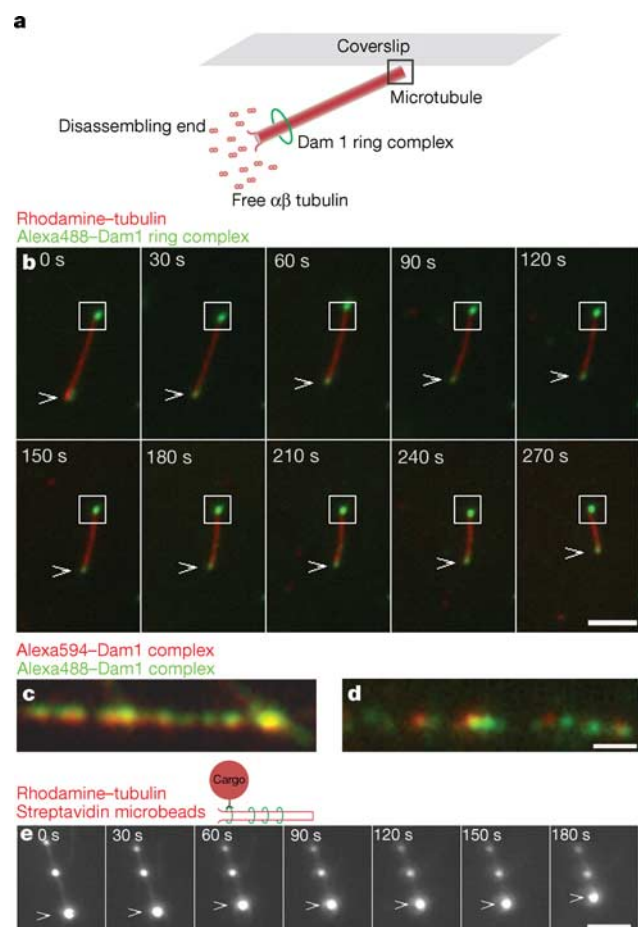


Figure 1 | A microscopy assay to study the interaction of the Dam1 complex with dynamic microtubules *in vitro*. **a**, Sketch of the assay. Observation of Dam1 ring (green) behaviour requires microtubules (red) that are attached to the coverslip with only one end or are completely free in solution. **b**, The Dam1 ring complex moves processively at the ends of depolymerizing microtubules. Time-lapse images of a microtubule (red) attached with one end to the coverslip (boxed area) while its free end depolymerizes after dilution. The Alexa488-Dam1 complex (green) follows the depolymerizing end (white arrowhead) without detaching from the lattice. Note that the fluorescent signal of the complex does not decrease significantly over the time of observation. Scale bar, 2 μm . See also Supplementary Movies S1, S2 and S3. **c**, Experiment to test for turnover of ring subunits. Green and red fluorescent Dam1 complexes are first mixed, incubated with unlabelled microtubules, and then visualized. Each spot on the microtubule shows co-localization of red and green ring segments. **d**, Green fluorescent complex is incubated with microtubules first, then red fluorescent complex is added. Spots on the microtubule corresponding to Dam1 rings are either red or green, demonstrating that there is no turnover or exchange of ring subunits. Scale bar, 1 μm . **e**, Coupling of a cargo to Dam1 rings. Time-lapse images of Dam1-coupled microbeads on depolymerizing microtubule ends (see also Supplementary Movie S4). Depolymerization was induced by adding XMCAK1. Scale bar, 5 μm .

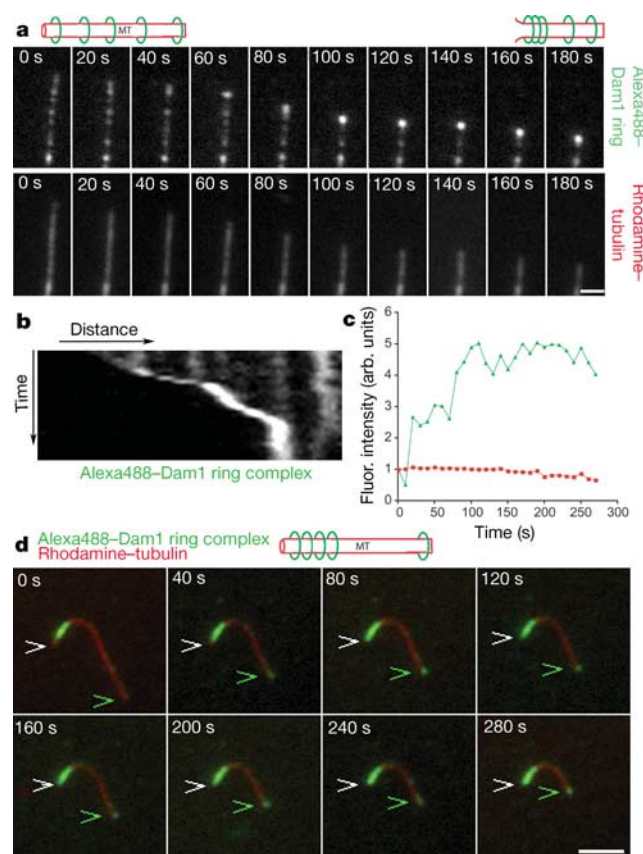


Figure 2 | Collection of multiple Dam1 rings at a depolymerizing microtubule end. **a**, Individual Dam1 rings are 'collected' by the depolymerizing microtubule (MT) end. Dam1 signals merge at the depolymerizing end and the fluorescence intensity increases. **b**, Kymograph of the Alexa488 signal from **a**, showing the behaviour of individual Dam1 rings. Note that individual lines merge, and then move together with increased intensity. **c**, Quantification of fluorescence intensity of the Alexa488 signal at the depolymerizing microtubule end. Note that the intensity does not increase linearly but in two major steps of equal size, corresponding to the merging of the strongest Dam1 signals. See also Supplementary Movie S5. **d**, Stabilization of depolymerizing microtubule ends depends on the Dam1 ring concentration: time-lapse images of a depolymerizing microtubule with asymmetric Dam1 decoration. Depolymerization of one end (white arrowhead) is stalled by a high concentration of Dam1 rings. The sparsely decorated end (green arrowhead) continues to depolymerize and the Dam1 ring follows the end without detaching. Depolymerization was induced by dilution. Scale bar, 2 μm . See also Supplementary Movie S6.

contrast, the intensity of the tubulin signal of the depolymerizing end remained constant over time (Fig. 2c). In many cases the collection of multiple Dam1 signals seemed to slow down or eventually stall depolymerization (Fig. 2b). For a better evaluation of the effect of Dam1 ring concentration on depolymerization rates we observed microtubules with asymmetrically distributed Dam1 rings: as shown in Fig. 2d (Supplementary Movie S6), a sparsely decorated end (green arrowhead) depolymerized over time and the Dam1 signal tracked with the end as observed previously. In contrast, the other end (white arrowhead) initially started to depolymerize (0–60 s) but then stalled at an area with a high Dam1 concentration (60–280 s). Thus, direct observation indicates that two fundamental biochemical activities of the complex, namely stabilization and mobility along the lattice, depend on the relative concentration of Dam1 rings at the microtubule ends.

We next investigated the dynamics of individual Dam1 rings on stable microtubules. Alexa488–Dam1 complex was incubated at a final concentration of 1–5 nM with Taxol-stabilized microtubules and followed by using fluorescence microscopy at high time resolution to detect small movements. On observation of the whole field, numerous Dam1 spots could be seen undergoing small random motions without directional preference (Supplementary Movie S7). On single Taxol-stabilized microtubules individual Dam1 spots moved independently of each other, changed direction, merged, and separated again (Fig. 3a and Supplementary Movie S8). In contrast, Dam1 signals that were adsorbed on the coverslip surface

remained immobile (Supplementary Movie S9). Thus, individual Dam1 rings show undirected, one-dimensional diffusion on a stable microtubule lattice. The mean-square displacement of individual spots increased linearly as a function of time (Fig. 3b). From the slope of the graph we calculated an average diffusion constant of about $0.54 \times 10^{-9} \text{ cm}^2 \text{ s}^{-1}$ for the movement, which is comparable to values reported for other microtubule–motor-protein systems^{7,20}.

The sliding of a ring complex along the lattice is a previously unobserved mode of microtubule-based motility and we sought to investigate structural determinants for such a mechanism. When we saturated Taxol-stabilized microtubules with Dam1 complex and revealed the sample by negative-stain electron microscopy, we occasionally observed ‘end-on views’ that provided axial projections of the microtubule–Dam1 complex arrangement (Fig. 4a). In the best three images the number of protofilaments (13) could clearly be counted, but the individual Dam1 complexes in the ring density around the microtubule were not as well defined. We performed rotational analysis on three end-on images after masking out the

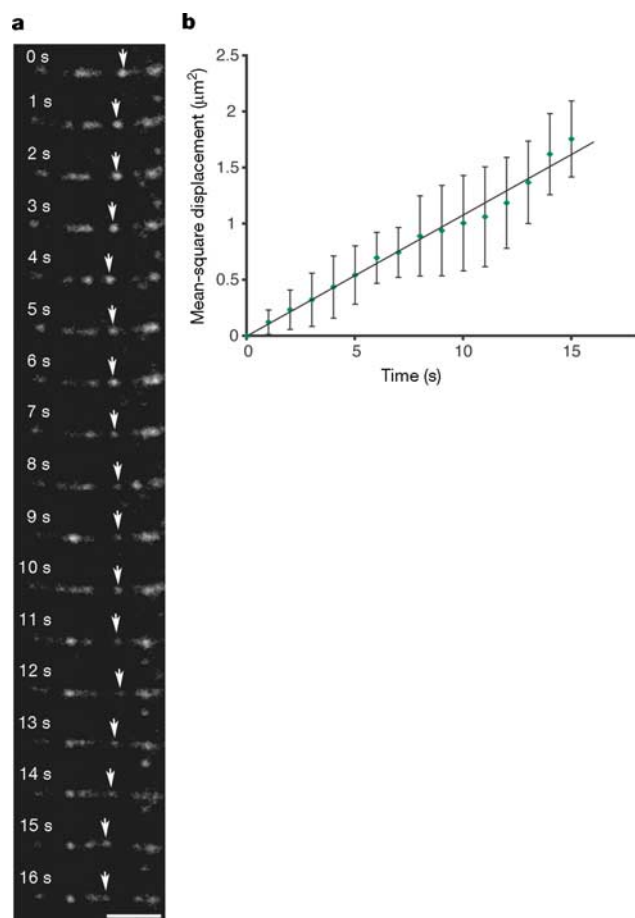


Figure 3 | One-dimensional diffusion of Dam1 rings on stable microtubules. **a**, An individual Dam1 spot on a Taxol-stabilized microtubule is followed over time. Note that it moves in both directions while other spots merge and separate again. Scale bar, 2 μm . **b**, Plot of mean-square displacement of individual Dam1 signals against time. Error bars show standard deviation. See also Supplementary Movies S7, S8 and S9.

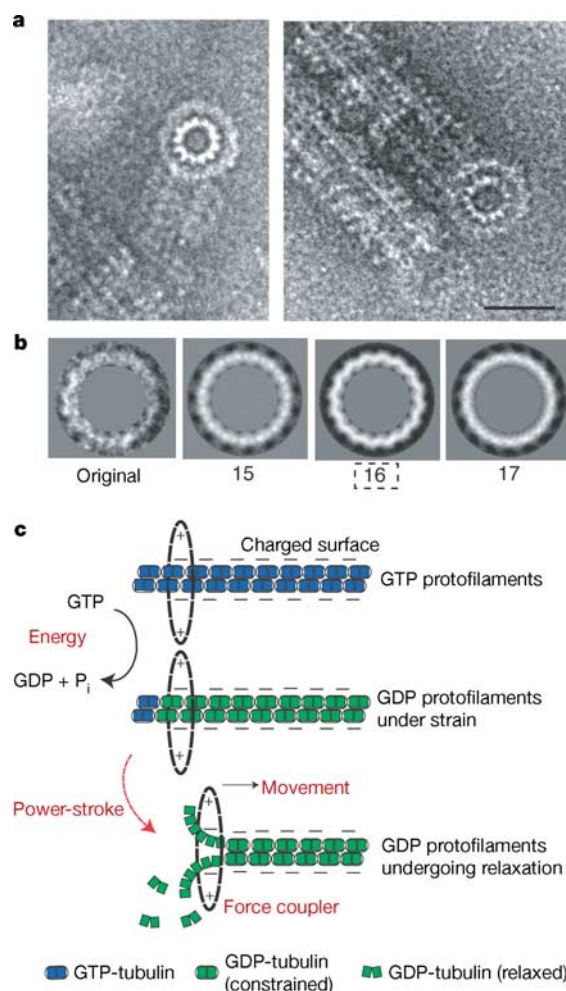


Figure 4 | Image analysis of the Dam1 ring complex around microtubules. **a**, Examples of electron micrographs showing ‘end-on views’ of microtubules surrounded with Dam1 complex. Scale bar, 50 nm. **b**, Rotational averaging of Dam1 complexes in ‘end-on views’. Original ring complex and reconstructions assuming different rotational symmetries are shown. A reconstruction with a 16-fold rotational symmetry shows sharper features and most closely resembles the original complexes. **c**, Model for Dam1 ring dynamics and function at the depolymerizing microtubule end. The free energy of GTP hydrolysis is released as peeling protofilaments produce a power-stroke by means of a transition from straight to curved. The Dam1 complex transduces this mechanical energy into directed sliding along the microtubule lattice to pull chromosomes apart during anaphase.

microtubule density. This analysis revealed a consistent 16-fold symmetry for the Dam1 ring. The sharpness of different rotational averages of the ring density at and around the correct rotational symmetry supports the cross-correlation analysis and provides a signal-enhanced view of the arrangement of the 16 Dam1 complexes as they form a ring around the microtubule (Fig. 4b).

The symmetry mismatch between the 13-fold microtubule and the 16-fold Dam1 ring suggests that there might not be specific interactions between the surface of the microtubule and the Dam1 ring. Other combinations of microtubules (14 protofilament microtubules assembled with guanosine-5'-[(α,β)-methylene]triphosphate (GMPCPP)) and Dam1 rings (18-fold for instance) were also observed with symmetry mismatch. These observations, and the requirement of the tubulin C termini for the interaction with and assembly of the Dam1 rings, are consistent with a sliding mechanism in which positive charges in the inner surface of the Dam1 ring interact electrostatically with the negative charges along the microtubule surface³.

Thus, we have obtained conclusive evidence that the Dam1 ring complex, a critical microtubule-binding element of the budding yeast kinetochore, diffuses along the microtubule lattice and moves processively with the end of a depolymerizing microtubule. This processive movement could help chromosomes to stay attached to disassembling kinetochore microtubules during anaphase A and contribute to the successful segregation of chromosomes in the absence of minus-end-directed motors. Typically, microtubules have been viewed as static tracks along which motor proteins transport cargo. However, microtubules themselves produce pushing and pulling forces that can move cellular structures²¹. Our analysis of Dam1 motility shows that the microtubule end and the ring complex form a molecular machine that generates movement (Fig. 4c). Because the GTP and GDP forms of microtubule protofilaments differ in their preferred conformations^{22,23}, the chemical energy of GTP hydrolysis is stored as mechanical strain in the lattice and is released as a power-stroke by the outward peeling of protofilaments^{24–26}. The ring is needed as a force-coupling device that translates the mechanical energy of the power-stroke into a sustained movement along the lattice. As the ring moves in response to the polymerization state of the microtubule end, we also note that it could be part of the tension-sensing machinery that monitors kinetochore–microtubule attachments²⁷.

Despite its remarkable properties *in vitro*, the Dam1 complex is clearly not the only factor that contributes to dynamic kinetochore attachment at microtubule ends *in vivo*. Some chromosome segregation occurs in *dam1* mutants at the restrictive temperature, and the fission yeast Dam1 complex, although also contributing to high-fidelity chromosome segregation, is not essential for viability^{28,29}. The Dam1 complex could be especially important in budding yeast because only one microtubule attaches to each kinetochore.

METHODS

In vitro assays. Alexa488-labelled or 594-labelled Dam1 complex was prepared as described, and electron microscopy showed regular ring formation of the complex³. In an analogous procedure, biotin-maleimide (Sigma) was coupled to Dam1 complex and free biotin was separated by gel filtration. Flow cells were constructed with two layers of double-sided tape and 12 mm $\times$ 12 mm coverslips to create a chamber volume of about 15 μ l. To block surfaces they were pretreated with 0.2% (w/v) Pluronic F108 (BASF) in PEM medium (80 mM PIPES/KOH pH 6.8, 1 mM MgCl₂, 1 mM EGTA). Motility reactions (20 μ l) contained 0.8% (w/v) methylcellulose (Sigma), 0.2% (w/v) Pluronic F108, 200 μ g ml⁻¹ glucose oxidase, 35 μ g ml⁻¹ catalase, 4.5 mg ml⁻¹ glucose, 0.5% 2-mercaptoethanol, 1 mg ml⁻¹ BSA and 0.5 mg ml⁻¹ casein in PEM and the indicated amounts of Alexa488–Dam1 complex and microtubules. To follow XMCAP1-catalysed depolymerization, 1 μ l of rhodamine-labelled GMPCPP microtubules (15 μ M) was incubated with 1 μ l of Alexa488–Dam1 complex (0.4 μ M) for 5 min at 37 °C. Then 1 μ l of the mixture was combined with 19 μ l of motility buffer containing 15 nM XMCAP1 and 1.5 mM MgATP. Alternatively, 30 μ M microtubules were assembled in the presence of 1 mM GTP and 25% (v/v) glycerol, and 1 μ l of these microtubules was then incubated with 1 μ l of Alexa488–Dam1 complex and the

sample was diluted 1:10 into motility buffer containing 8 μ M unlabelled tubulin. The sample was introduced into the flow cell and observed immediately at 25 °C on an inverted Olympus IX 81 epi-illuminated microscope with a 100 $\times$ objective (numerical aperture 1.4) and neutral-density filters. For the observation of processive movements on depolymerizing microtubule ends, two-channel movies were made with a green fluorescent protein/red fluorescent protein filter set and motorized excitation and emission filter-wheels. The camera and the filter-wheels were controlled by Metamorph software (Universal Imaging). General processing of movies and images was performed with the ImageJ software (rsb.info.nih.gov/ij/).

Electron microscopy and symmetry analysis. Microtubule–Dam1 complexes were revealed by electron microscopy of negatively stained samples. Dam1 complex (20 μ l, 30 μ M) was incubated with 20 μ l of Taxol-stabilized microtubules (3.75 μ M) for 15 min at 37 °C, pelleted at 139,000g in a TL100 rotor and resuspended in 40 μ l of PEM (containing 1 mM GTP and 10 μ M Taxol). The sample was stained with uranyl acetate and electron microscopic images were recorded as described previously³. The rotational symmetry of the Dam1 complexes around the microtubules as seen in end-on views was analysed with the SPIDER image processing package³⁰. The end-on microtubule–Dam1 ring image was translated so that its rotation centre coincided with the image centre. The microtubule density was masked so that only the outer density corresponding to the Dam1 ring remained. The ring was then rotated at different angles and the cross-correlation coefficients with the original ring were calculated. A maximum in the cross-correlation profile as a function of the rotation angle was found every 22.5°, corresponding to a 16-fold rotational symmetry. Different rotational averages of the Dam1 ring density at and around the correct rotational symmetry (16-fold) were then calculated with SPIDER.

Received 22 September; accepted 9 November 2005.

Published online 15 January 2006.

1. Maiato, H., Deluca, J., Salmon, E. D. & Earnshaw, W. C. The dynamic kinetochore–microtubule interface. *J. Cell Sci.* **117**, 5461–5477 (2004).
2. Cheeseman, I. M. *et al.* Implication of a novel multiprotein Dam1p complex in outer kinetochore function. *J. Cell Biol.* **155**, 1137–1145 (2001).
3. Westermann, S. *et al.* Formation of a dynamic kinetochore–microtubule interface through assembly of the Dam1 ring complex. *Mol. Cell* **17**, 277–290 (2005).
4. Miranda, J. J., De Wulf, P., Sorger, P. K. & Harrison, S. C. The yeast DASH complex forms closed rings on microtubules. *Nature Struct. Mol. Biol.* **12**, 138–143 (2005).
5. Inoue, S. & Salmon, E. D. Force generation by microtubule assembly/disassembly in mitosis and related movements. *Mol. Biol. Cell* **6**, 1619–1640 (1995).
6. Koshland, D. E., Mitchison, T. J. & Kirschner, M. W. Polewards chromosome movement driven by microtubule depolymerization *in vitro*. *Nature* **331**, 499–504 (1988).
7. Lombillo, V. A., Stewart, R. J. & McIntosh, J. R. Minus-end-directed motion of kinesin-coated microspheres driven by microtubule depolymerization. *Nature* **373**, 161–164 (1995).
8. Coue, M., Lombillo, V. A. & McIntosh, J. R. Microtubule depolymerization promotes particle and chromosome movement *in vitro*. *J. Cell Biol.* **112**, 1165–1175 (1991).
9. Hill, T. L. Theoretical problems related to the attachment of microtubules to kinetochores. *Proc. Natl Acad. Sci. USA* **82**, 4404–4408 (1985).
10. Molodtsov, M. I., Grishchuk, E. L., Efremov, A. K., McIntosh, J. R. & Ataullakhanov, F. I. Force production by depolymerizing microtubules: a theoretical study. *Proc. Natl Acad. Sci. USA* **102**, 4353–4358 (2005).
11. Straight, A. F., Marshall, W. F., Sedat, J. W. & Murray, A. W. Mitosis in living budding yeast: anaphase A but no metaphase plate. *Science* **277**, 574–578 (1997).
12. Maddox, P. S., Bloom, K. S. & Salmon, E. D. The polarity and dynamics of microtubule assembly in the budding yeast *Saccharomyces cerevisiae*. *Nature Cell Biol.* **2**, 36–41 (2000).
13. De Wulf, P., McAnish, A. D. & Sorger, P. K. Hierarchical assembly of the budding yeast kinetochore from multiple subcomplexes. *Genes Dev.* **17**, 2902–2921 (2003).
14. Westermann, S. *et al.* Architecture of the budding yeast kinetochore reveals a conserved molecular core. *J. Cell Biol.* **163**, 215–222 (2003).
15. Cheeseman, I. M., Drubin, D. G. & Barnes, G. Simple centromere, complex kinetochore: linking spindle microtubules and centromeric DNA in budding yeast. *J. Cell Biol.* **157**, 199–203 (2002).
16. Cheeseman, I. M., Enquist-Newman, M., Muller-Reichert, T., Drubin, D. G. & Barnes, G. Mitotic spindle integrity and kinetochore function linked by the Duo1p/Dam1p complex. *J. Cell Biol.* **152**, 197–212 (2001).
17. Cheeseman, I. M. *et al.* Phospho-regulation of kinetochore-microtubule attachments by the Aurora kinase Iplp. *Cell* **111**, 163–172 (2002).
18. Tanaka, K. *et al.* Molecular mechanisms of kinetochore capture by spindle microtubules. *Nature* **434**, 987–994 (2005).

19. Kapitein, L. C. *et al.* The bipolar mitotic kinesin Eg5 moves on both microtubules that it crosslinks. *Nature* **435**, 114–118 (2005).
20. Vale, R. D., Soll, D. R. & Gibbons, I. R. One-dimensional diffusion of microtubules bound to flagellar dynein. *Cell* **59**, 915–925 (1989).
21. Howard, J. & Hyman, A. A. Dynamics and mechanics of the microtubule plus end. *Nature* **422**, 753–758 (2003).
22. Wang, H. W. & Nogales, E. Nucleotide-dependent bending flexibility of tubulin regulates microtubule assembly. *Nature* **435**, 911–915 (2005).
23. Hyman, A. A., Chretien, D., Arnal, I. & Wade, R. H. Structural changes accompanying GTP hydrolysis in microtubules: information from a slowly hydrolyzable analogue guanylyl-(α,β)-methylene-diphosphonate. *J. Cell Biol.* **128**, 117–125 (1995).
24. Mandelkow, E. M., Mandelkow, E. & Milligan, R. A. Microtubule dynamics and microtubule caps: a time-resolved cryo-electron microscopy study. *J. Cell Biol.* **114**, 977–991 (1991).
25. Mahadevan, L. & Mitchison, T. J. Cell biology: powerful curves. *Nature* **435**, 895–897 (2005).
26. Grishchuk, E. L., Molodtsov, M. I., Ataulakhov, F. I. & McIntosh, J. R. Force production by disassembling microtubules. *Nature* **438**, 384–388 (2005).
27. Pinsky, B. A. & Biggins, S. The spindle checkpoint: tension versus attachment. *Trends Cell Biol.* **15**, 486–493 (2005).
28. Liu, X., McLeod, I., Anderson, S., Yates, J. R. & He, X. Molecular analysis of kinetochore architecture in fission yeast. *EMBO J.* **24**, 2919–2930 (2005).
29. Sanchez-Perez, I. *et al.* The DASH complex and Klp5/Klp6 kinesin coordinate bipolar chromosome attachment in fission yeast. *EMBO J.* **24**, 2931–2943 (2005).
30. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Walczak for the gift of recombinant XMCAK1; M. Kaksonen and C. Toret for help with image analysis; and all members of the Barnes/Drubin laboratory for discussions. This work was supported by grants from the National Institute of General Medical Sciences to G.B. and E.N., and from the Office of Basic Energy Science of the US Department of Energy. E.N. is a Howard Hughes Medical Institute Investigator. Research in this study was in part supported by a grant to D.G.D. from Phillip Morris USA Inc. and Phillip Morris International and by a postdoctoral fellowship of the Deutsche Forschungsgemeinschaft to S.W.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.B. (gbarnes@socrates.berkeley.edu).

LETTERS

Atomic structure of a Na⁺- and K⁺-conducting channelNing Shi¹, Sheng Ye¹, Amer Alam¹, Liping Chen¹ & Youxing Jiang¹

Ion selectivity is one of the basic properties that define an ion channel. Most tetrameric cation channels, which include the K⁺, Ca²⁺, Na⁺ and cyclic nucleotide-gated channels, probably share a similar overall architecture in their ion-conduction pore, but the structural details that determine ion selection are different. Although K⁺ channel selectivity has been well studied from a structural perspective^{1,2}, little is known about the structure of other cation channels. Here we present crystal structures of the NaK channel from *Bacillus cereus*, a non-selective tetrameric cation channel, in its Na⁺- and K⁺-bound states at 2.4 Å and 2.8 Å resolution, respectively. The NaK channel shares high sequence homology and a similar overall structure with the bacterial KcsA K⁺ channel, but its selectivity filter adopts a different architecture. Unlike a K⁺ channel selectivity filter, which contains four equivalent K⁺-binding sites, the selectivity filter of the NaK channel preserves the two cation-binding sites equivalent to sites 3 and 4 of a K⁺ channel, whereas the region corresponding to sites 1 and 2 of a K⁺ channel becomes a vestibule in which ions can diffuse but not bind specifically. Functional analysis using an ⁸⁶Rb flux assay shows that the NaK channel can conduct both Na⁺ and K⁺ ions. We conclude that the sequence of the NaK selectivity filter resembles that of a cyclic nucleotide-gated channel and its structure may represent that of a cyclic nucleotide-gated channel pore.

All K⁺ channels contain the highly conserved signature sequence TVGYG, which is essential for K⁺ ion selectivity^{3,4}. The cyclic nucleotide-gated (CNG) channel pore shares high sequence homology with K⁺ channels, but is non-selective and permeable to most group IA monovalent cations^{5–8}. A key difference between the pores of CNG and K⁺ channels is that CNG channels lack a tyrosine and glycine residue from the conserved TVGYG sequence (bold) present in K⁺ channels. This causes structural differences in the selectivity filter and results in the loss of ion selectivity in CNG channels.

In a search of the microbial genome, we identified two two-transmembrane channels from *Bacillus cereus* and *Bacillus anthracis* that have sequences very similar to the KcsA K⁺ channel, except for their selectivity filters, which resemble those of CNG channels with the sequence TVGDG or TVGDA (Supplementary Fig. S1). On the basis of this similarity, we hypothesized that these channels can non-selectively conduct both Na⁺ and K⁺. This proved to be correct according to our functional assay. Here we present the crystal structure of the channel from *Bacillus cereus* in complex with Na⁺ at 2.4 Å (Table 1 and Supplementary Table S1), and we call this channel NaK (conducting both Na⁺ and K⁺ ions).

NaK shares several common features with KcsA (Fig. 1a, b). It contains two membrane-spanning segments, M1 and M2, corresponding to the 'outer' and 'inner' helices of KcsA, and the four inner helices form an 'inverted teepee'. The ion conduction pathway has a water-filled cavity near the centre of the membrane, with four pore

helices pointing their carboxy termini towards the cavity. This overall architecture shared by NaK and KcsA presumably underlies their function to conduct cations across the cell membrane.

Two notable differences distinguish NaK and KcsA from each other. First, the amino-terminal 19 amino acids of NaK form an interfacial helix (M0 helix) parallel to the membrane. Phe 4, Leu 8 and Met 11 at the N-terminal region of each M0 helix form hydrophobic contacts with Val 26, Val 29 and Leu 30 at the N terminus of helix M1 from a neighbouring subunit (Fig. 1b). The four M0 helices form a cuff that encircles the inner helix bundle crossing, and seem to be positioned appropriately to affect the opening and closing of the pore.

The second structural difference between NaK and KcsA is observed in the selectivity filter (Fig. 2). The KcsA filter is formed by four extended polypeptide chains, each containing the ₇₅TVGYG₇₉ sequence (Fig. 2b, left). The backbone carbonyl oxygen atoms from the TVGY residues, along with the hydroxyl oxygen atom from Thr 75, point towards the centre of the pathway and form four equivalent binding sites (numbered 1 to 4) for dehydrated K⁺ ions. The eight oxygen atoms surrounding each site mimic the hydration shell of a K⁺ ion. Four additional carbonyl oxygen atoms from Gly 79 sit at the perimeter of the filter entrance and point directly into the extracellular solution, generating an electronegative environment that stabilizes a half-dehydrated K⁺ ion. The side chains from the four Tyr 78 residues form specific packing interactions with neighbouring aromatic residues from the pore helices that surround the filter.

In comparison, the NaK filter has a sequence of ₆₃TVGDG₆₇, with the hydroxyl oxygen atom from Thr 63 and backbone carbonyl oxygen atoms from Thr 63 and Val 64 forming two ion-binding sites equivalent to sites 3 and 4 in KcsA (Fig. 2b, right). For comparative purposes, these two sites are also numbered 3 and 4 in Fig. 2b. Sites 1 and 2 of KcsA do not exist in NaK owing to the

Table 1 | Refinement statistics

	Na ⁺ complex	K ⁺ complex
Resolution (Å)	30–2.4	30–2.8
R _{work} /R _{free} (%)	23.5/26.0	24.1/28.0
Number of atoms		
Protein	1,648	1,655
Ligand/ion	5	5
Water	20	15
B-factors		
Protein	73.74	72.85
Ligand/ion	73.45	72.77
Water	52.31	73.75
r.m.s deviations		
Bond lengths (Å)	80.97	72.87
Bond angles (°)	0.006	0.007
	1.08	1.16

¹Department of Physiology, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9040, USA.

different backbone conformations adopted by the GDG residues at these two positions. The carbonyl groups from Gly65 and Asp66 align tangentially to the ion conduction pathway, and therefore their oxygen atoms do not point towards the centre to coordinate ions. Furthermore, the C α of Asp66 is directed away from the centre and its side chain protrudes upward and is exposed to the extracellular solution, creating a vestibule where ions can diffuse but not bind very specifically. At the entryway to the filter from the extracellular solution, the main chain at Gly67 pinches inwards so that its carbonyl oxygen points towards the pore axis, forming an ion-binding site, which, as we discuss below, can have important physiological consequences.

Superimposition of the NaK filter with that of KcsA gives rise to a root-mean-square deviation (r.m.s.d.) of 1.6 Å in their main-chain positions. These differences are caused mainly by the last three residues, which have an r.m.s.d. of 2.1 Å. The observed differences between NaK and KcsA are not due to crystal packing, as the NaK filter is in the centre of the channel tetramer and is not involved in protein packing in the crystal.

In KcsA, K⁺ ions are required to stabilize the selectivity filter: the structure of KcsA in conditions of high Na⁺, low K⁺ reveals significant changes at the filter region⁹. In contrast, the structure of NaK does not depend on the ion composition: we determined a structure in KCl (rather than NaCl) at 2.8 Å and found no apparent differences in the filter structure (Fig. 3a–c). Thus, NaK retains its proper conformation for ion conduction in both Na⁺ and K⁺.

Electron density maps of NaK show that ions can bind at the extracellular entrance, along the selectivity filter, and in the central cavity of the ion conduction pore (Fig. 2a). Ion-omit maps calculated with data from crystals grown in NaCl (Fig. 3a) or KCl (Fig. 3b) show that these ions bind at all the same sites in the pore. The crystallization solutions also contained 200 mM CaCl₂ in addition to 100 mM NaCl or KCl. To deduce whether it is the monovalent or divalent cations that bind at specific sites within the pore, we carried out soaking experiments in which the NaK crystals were soaked in stabilization solutions containing various monovalent and divalent salts at the same concentrations as those in the crystallization solutions. To minimize non-isomorphism, a native reference crystal was subjected to the same soaking procedures in a stabilization solution containing CaCl₂ and NaCl (Supplementary Table S2). Difference electron density maps using Fourier coefficients ($F_{\text{soak}} - F_{\text{reference}}$) were calculated with phases from the model (Fig. 3d–h).

The electron density difference between crystals soaked with Na/Ca and Cs/Ca, Na/Ca and Rb/Ca, or Na/Ca and Tl/Ca all give rise to three strong peaks at sites 3, 4 and the central cavity, indicating

the binding of Cs⁺, Rb⁺ and Tl⁺ at these sites (Fig. 3d–f, red mesh). The difference map between Na/Ca- and Na/Mg-soaked crystals reveals only one Ca²⁺-binding site, at the extracellular entrance to the filter (Fig. 3g, green mesh). This suggests that only monovalent cations bind at sites 3, 4 and the central cavity, and that divalent cations can bind at the external site outside the selectivity filter. On the basis of these results, looking back at the ion omit maps we can conclude that it is Na⁺ and K⁺, but not Ca²⁺, inside the selectivity filter and central cavity (Fig. 3a, b).

The Ca²⁺-binding site formed by the four carbonyl oxygen atoms from the Gly67 residues may not be specific for divalent cations—Ca²⁺ binding might have prevailed simply because it is at a higher concentration in the crystallization solutions. A $2F_o - F_c$ map of a native crystal soaked in a stabilizing solution containing only NaCl still shows reasonably strong electron density at this position (data not shown), possibly owing to Na⁺ binding. Monovalent cation binding at this position probably occurs as ions conduct through the pore. Divalent cations might bind at this position to block the conduction of monovalent cations through the pore, a property relevant to CNG channels, for which a divalent cation block of monovalent currents is important for physiological channel function^{10–15}.

The difference map between Na/Ca and Na/Ba soaked crystals reveals two Ba²⁺-binding sites, one at the filter entrance (equivalent to the Ca²⁺-binding site) and another at site 3 in the selectivity filter (Fig. 3h, green mesh). Ba²⁺ can block a K⁺ channel^{16,17} by binding to site 4 in the selectivity filter¹⁸. Ba²⁺ might also serve as a blocker for the NaK channel by binding either to site 3 (instead of site 4 as it does in KcsA) of the selectivity filter or at the external entrance. Indeed, our functional assay showed that Ba²⁺ reduces the Rb⁺ flux at micromolar concentrations (data not shown).

Neither the ion-omit maps of native crystals nor the difference maps between various soaked crystals reveal any specific ion binding in the vestibule of the filter. The $2F_o - F_c$ maps of native crystals (Fig. 2a) and the difference map between Na/Ca- and Tl/Ca-soaked crystals reveal weak electron density in the vestibule, indicating the presence of an ion with low occupancy. Ion binding in the vestibule can be stabilized by the electronegative environment generated by the surface lining main-chain carbonyl groups from Gly65 and Asp66. However, such binding is not as specific as that of sites 3 and 4, and the bound ion distributes over a larger volume, giving rise to weak electron density on both $2F_o - F_c$ and difference maps.

Based on the premise that NaK is capable of conducting Rb⁺ (among other monovalent cations), we performed an ⁸⁶Rb flux assay¹⁹ using NaK channels reconstituted into liposomes in order

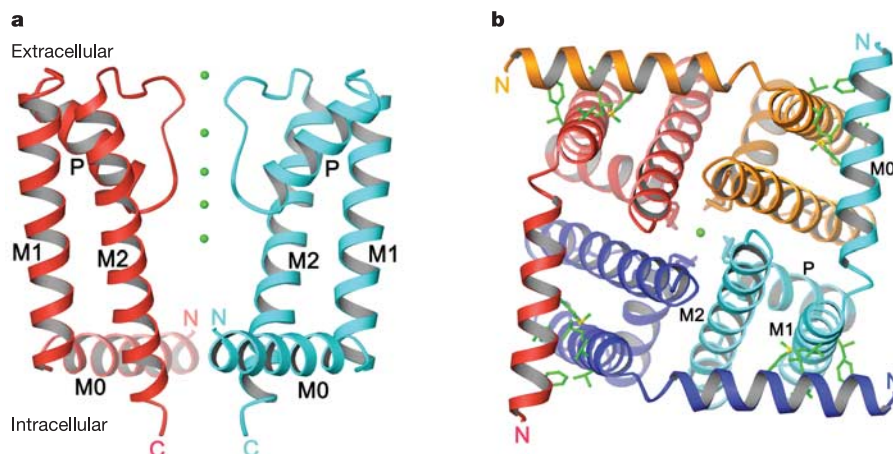


Figure 1 | Overall structure of NaK. **a**, **b**, Ribbon representation of NaK viewed from the membrane in **a** (with front and rear subunits removed), and from the intracellular side in **b**. Green spheres represent ions in the channel.

Each subunit is individually coloured. Green ball-and-stick representations show side chains from residues involved in hydrophobic contacts between M0 and neighbouring M1 helices.

to study channel ion-selectivity properties. We also reconstituted a truncated form of the channel, NaKN Δ 19, in which the N-terminal M0 helix-forming residues were removed, as we suspected that the four M0 helices around the helix bundle might lock the channel in a closed form. A Na $^{+}$ concentration gradient was established across the liposome membrane (high concentration of NaCl inside) so that any outflow of Na $^{+}$ through NaK would leave a deficit of positive charge inside and drive the influx of external radioactive 86 Rb, which could be monitored using a scintillation counter. Figure 4a shows the time-dependent accumulation of 86 Rb inside liposomes, confirming that NaK is capable of conducting both Na $^{+}$ and Rb $^{+}$. As a much higher flux rate was observed for NaKN Δ 19, this truncated form of NaK was used in all other flux assays. As expected, the flux rate was negligible for control liposomes (containing no reconstituted protein) and for liposomes reconstituted with KcsA, which is a K $^{+}$ -selective channel. Addition of gramicidin A, a Group 1A metal ionophore, to the control liposomes increases Rb $^{+}$ influx, confirming that it is the outflow of Na $^{+}$ that causes Rb $^{+}$ influx.

We performed the same flux assay using liposomes loaded with KCl instead of NaCl. As shown in Fig. 4b, liposomes reconstituted with KcsA or NaKN Δ 19 both show 86 Rb accumulation in a time-dependent manner, indicating the conduction of K $^{+}$ in both channels. The difference in flux rates of NaKN Δ 19 in the presence of Na $^{+}$ or K $^{+}$ could be the result of different batches of liposomes used for each set of experiments.

We then measured 86 Rb influx in the presence of external cations at different concentrations in addition to 86 Rb. NaKN Δ 19-containing liposomes were loaded with NaCl and flux was allowed to proceed for 10 min before radioactivity levels in the liposomes were measured. Any of the test ions that the channel was able to conduct were expected to compete with 86 Rb and decrease its influx into liposomes. The data for these 'competition assays' are shown in Fig. 4c and indicate that the channel is able to conduct both Na $^{+}$ and K $^{+}$, which result in lowering of the 86 Rb signal compared to control (86 Rb only, in a background of $\sim 35 \mu\text{M}$ NaCl). Li $^{+}$ and NMG $^{+}$ (N-methyl-D-glucamine) do not lower the 86 Rb signal and hence we conclude that the channel is unable to conduct these cations. Competition assays repeated in liposomes with KCl yielded the same pattern of results (data not shown).

From the competition assays, it seems that K $^{+}$ has a greater effect on 86 Rb flux than Na $^{+}$, and that the channel might therefore conduct K $^{+}$ ions better than Na $^{+}$ ions. However, we are reluctant to draw such conclusions before a more thorough and quantitative functional analysis of the channel is performed. What is clear from the data is that the NaK channel is not very selective among Na $^{+}$, K $^{+}$ and Rb $^{+}$. This result is consistent with our crystallographic studies showing that these ions all bind similarly inside the selectivity filter and central cavity of the NaK channel.

One of the more intriguing conclusions we draw from the NaK channel has to do with ion-binding sites 3 and 4. Chemically, these

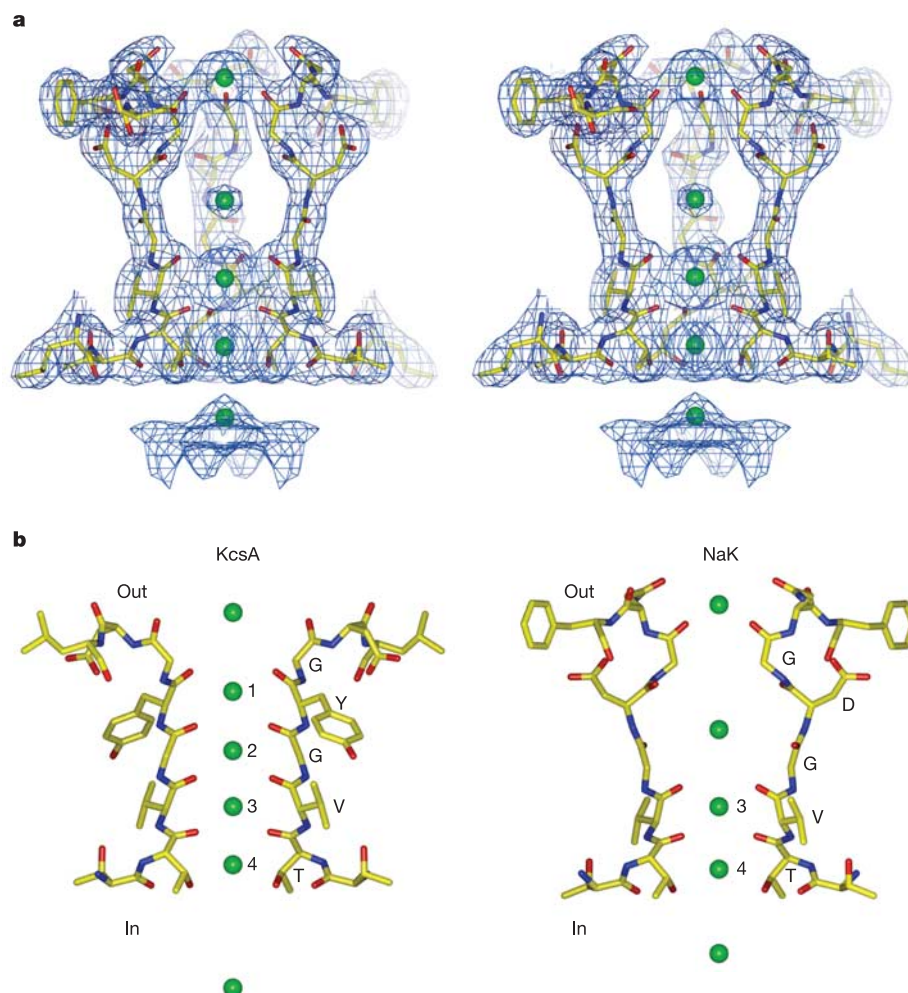


Figure 2 | Structural comparison of the selectivity filters in NaK and KcsA. **a**, Stereo view of a $2F_o - F_c$ map at 1σ contour (blue mesh), showing the selectivity region of NaK with the front subunit removed. Green spheres represent ions in the filter and cavity. **b**, Structural details of the selectivity

filters from KcsA (left) and NaK (right). The front and rear subunits have been removed for clarity. Only discrete ion-binding sites in the filter are numbered (1–4 from the extracellular side in KcsA, 3 and 4 in NaK).

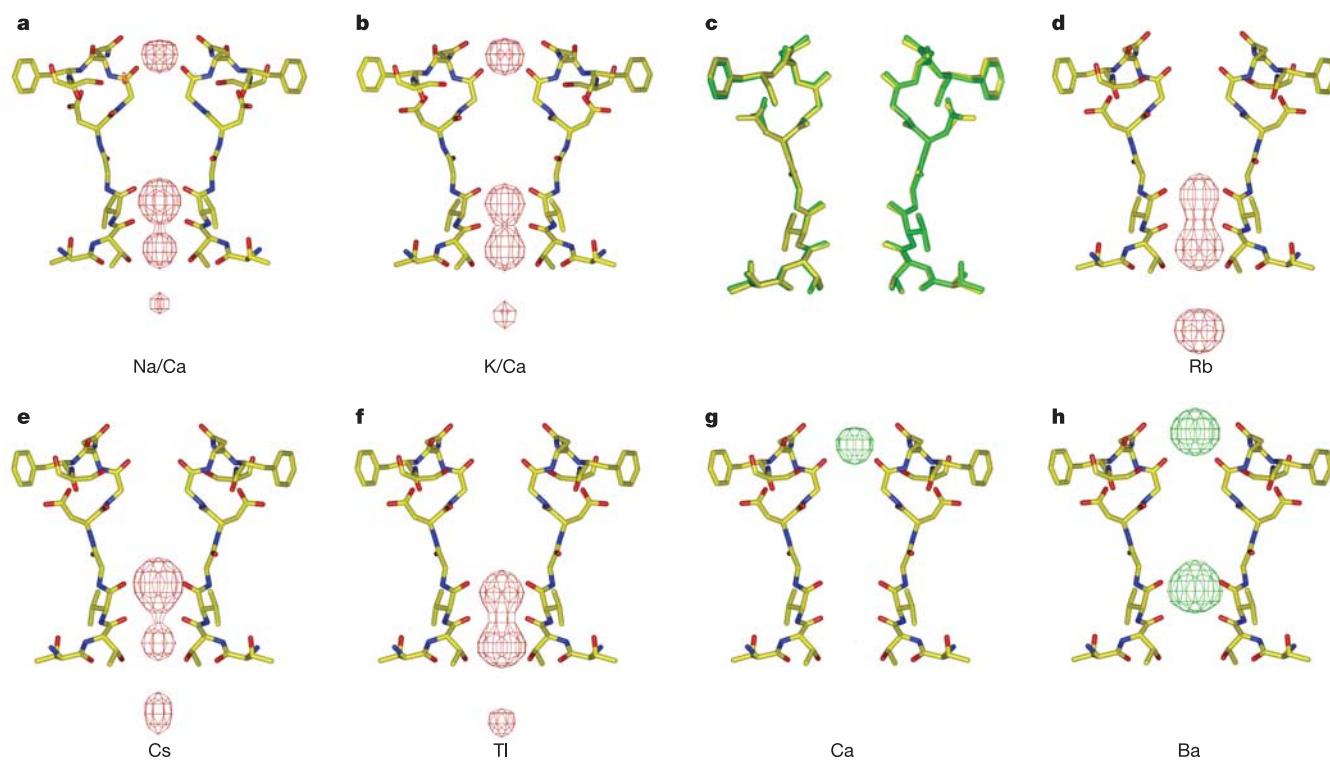


Figure 3 | Ion binding in the NaK channel. **a, b**, 2F_o-F_c ion-omit maps of Na⁺ (**a**) and K⁺ (**b**) complexes of NaK contoured at 6σ (red mesh). **c**, Superimposition of NaK selectivity filters in Na⁺- (green) and K⁺- (yellow) bound states. **d-h**, F_{soak}-F_{reference} difference maps between the reference crystal and crystals with various soaking conditions reveal the

binding of cations (labelled underneath each panel) in NaK. All maps are contoured at 10σ except for the density of Ca²⁺ binding, which is contoured at 6σ. Densities for monovalent and divalent cations are coloured red and green, respectively.

sites are indistinguishable from the corresponding positions in the KcsA K⁺ channel, yet their ion-binding properties are clearly different from those observed in K⁺ channels. Specifically, in K⁺ channels, Na⁺ is never observed to bind at position 3, whereas in the NaK channel, Na⁺ binds there without difficulty. On the basis of these experimental data, we conclude that structural differences in otherwise chemically identical sites must account for their different ion-binding properties. These structural differences must arise from differences in amino acid packing around the binding site.

Our structural study of NaK also addresses a fundamental issue concerning ion selectivity in K⁺ channels. A theory-based computational assay has suggested that K⁺ selectivity in K⁺ channels originates not from geometric constraints provided by the protein, but rather from electrostatic repulsion between carbonyl oxygen atoms—the repulsion was said to prevent the oxygen atoms from approaching close enough to each other to form a cage small enough

to coordinate the smaller Na⁺ ion²⁰. Comparison of NaK and KcsA provides a clear experimental demonstration that electrostatic repulsion between carbonyl oxygen atoms is not the origin of K⁺ over Na⁺ selectivity in K⁺ channels, and that protein atoms surrounding the ion-binding site must confer size-selectivity through geometric constraints.

We also find that the external site in the NaK selectivity filter is distinct from the ion-binding site just outside the selectivity filter of K⁺ channels (Fig. 2b). In K⁺ channels, the external site does not bind divalent cations, whereas in NaK it can bind divalent cations that could serve to block monovalent cation conduction. This difference can be understood on the basis of their different structural properties. In the K⁺ channel this region of the selectivity filter appears to be very constrained. An apparent greater flexibility of the main chain in this region of the NaK selectivity filter probably allows the carbonyl oxygen atoms to conform to an ion with more degrees of freedom.

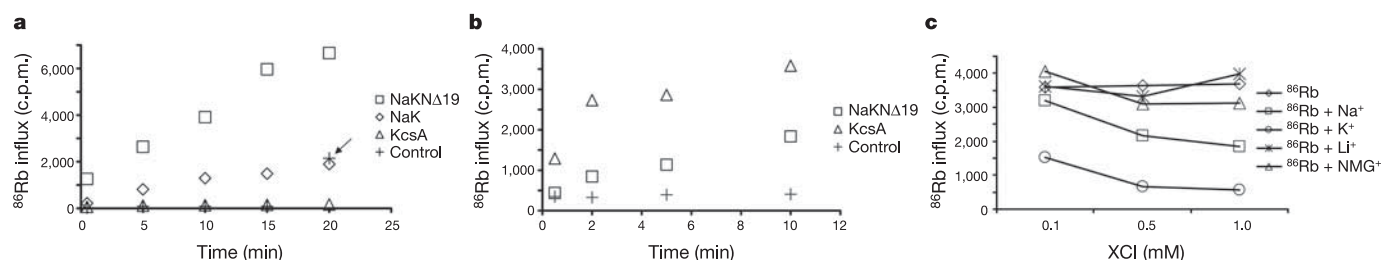


Figure 4 | ⁸⁶Rb flux assay. **a**, Time-dependent ⁸⁶Rb influx into liposomes prepared in NaCl. The liposomes contain NaK, NaKΔ19, KcsA or no protein (as a control). Arrow indicates ⁸⁶Rb influx into control liposomes upon addition of 10 μg ml⁻¹ gramicidin A. **b**, Time-dependent ⁸⁶Rb influx into KcsA or NaKΔ19-containing liposomes prepared in KCl.

c, Competition assay showing ⁸⁶Rb influx in the presence of externally added monovalent cations. Three final concentrations of each cation (0.1, 0.5 and 1.0 mM) were tested in the assay. No external cations were added in the control experiment. c.p.m., counts per minute.

Divalent cation binding at the external site in the NaK selectivity filter reveals the structural basis of external divalent cation blockage in CNG channels. The physiological significance of this blockage to visual transduction was first revealed nearly 20 years ago²¹. Here for the first time we observe its chemical basis in the form of a precise binding site at the entryway.

METHODS

Detailed experimental methods are provided in the Supplementary Information.

The NaK channel from *Bacillus cereus* was cloned into the pQE60 vector and expressed in *E. coli* XL1-Blue cultures. The protein was purified as a tetramer in *n*-decyl- β -D-maltoside with NaCl or KCl as the monovalent salt. Crystals were grown by sitting-drop vapour diffusion at 20 °C by mixing equal volumes of protein solution at 30–35 mg ml⁻¹ and reservoir solution containing 36–42% polyethylene glycol 400 (PEG400), 200 mM CaCl₂, 100 mM Tris HCl pH 7.5–8.5, 4% *t*-butanol. The crystals were of space group C222₁ with cell dimensions $a = 81.5$ Å, $b = 85.5$ Å, $c = 129.6$ Å, $\alpha = \beta = \gamma = 90^\circ$, and contained two subunits per asymmetric unit. The four-fold axis of the channel tetramer coincides with one of the crystallographic dyads.

All data were collected at the Advanced Photon Source (APS) and processed using HKL2000²². The structure was determined by single isomorphous replacement with anomalous scattering (SIRAS) using a mercury derivative. Hg binding sites were determined with SHELXD²³ and initial phases were improved by solvent flattening in *n*-decyl- β -D-maltoside²⁴. The model was constructed in O²⁵ and was refined in CNS²⁶ to 2.4 Å with $R_{\text{work}} = 23.5\%$ and $R_{\text{free}} = 26.0\%$, and contained residues 1–103 in one subunit, 1–104 in the other. The K⁺-bound NaK structure was determined by molecular replacement and was refined to 2.8 Å with $R_{\text{work}} = 24.1\%$ and $R_{\text{free}} = 28.0\%$.

For soaking experiments, crystals of Na⁺ complexes were soaked in a stabilization solution of 40% PEG400, 100 mM Tris HCl pH 8.0, 20 mM *n*-decyl- β -D-maltoside, 4% *t*-butanol, 100 mM XCl and 200 mM YCl₂, where X and Y represent a monovalent cation (Na⁺, Rb⁺ or Cs⁺) and a divalent cation (Ca²⁺, Mg²⁺ or Ba²⁺), respectively. For Tl⁺-soaking, 100 mM TlNO₃ and 200 mM Ca(NO₃)₂ were used.

All channel proteins used in the flux assay were reconstituted into lipid vesicles composed of 1-palmitoyl-2-oleoyl-phosphatidylethanolamine and 1-palmitoyl-2-oleoyl-phosphatidylglycerol using the same method as previously described²⁷. The ⁸⁶Rb flux assay was performed as described¹⁹. For competition assays, the tested ions (Li⁺, Na⁺, K⁺ or NMg⁺) were added directly into the flux buffer.

Received 28 July; accepted 5 December 2005.

Published online 8 February 2006.

1. Zhou, Y., Morais-Cabral, J. H., Kaufman, A. & MacKinnon, R. Chemistry of ion coordination and hydration revealed by a K⁺ channel–Fab complex at 2.0 Å resolution. *Nature* **414**, 43–48 (2001).
2. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* **280**, 69–77 (1998).
3. Heginbotham, L., Abramson, T. & MacKinnon, R. A functional connection between the pores of distantly related ion channels as revealed by mutant K⁺ channels. *Science* **258**, 1152–1155 (1992).
4. Heginbotham, L., Lu, Z., Abramson, T. & MacKinnon, R. Mutations in the K⁺ channel signature sequence. *Biophys. J.* **66**, 1061–1067 (1994).
5. Yau, K. W. & Baylor, D. A. Cyclic GMP-activated conductance of retinal photoreceptor cells. *Annu. Rev. Neurosci.* **12**, 289–327 (1989).
6. Kaupp, U. B. & Seifert, R. Cyclic nucleotide-gated ion channels. *Physiol. Rev.* **82**, 769–824 (2002).
7. Matulef, K. & Zagotta, W. N. Cyclic nucleotide-gated ion channels. *Annu. Rev. Cell Dev. Biol.* **19**, 23–44 (2003).
8. Zagotta, W. N. & Siegelbaum, S. A. Structure and function of cyclic nucleotide-gated channels. *Annu. Rev. Neurosci.* **19**, 235–263 (1996).
9. Zhou, Y. & MacKinnon, R. The occupancy of ions in the K⁺ selectivity filter: charge balance and coupling of ion binding to a protein conformational

change underlie high conduction rates. *J. Mol. Biol.* **333**, 965–975 (2003).

10. Haynes, L. W., Kay, A. R. & Yau, K. W. Single cyclic GMP-activated channel activity in excised patches of rod outer segment membrane. *Nature* **321**, 66–70 (1986).
11. Stern, J. H., Knutsson, H. & MacLeish, P. R. Divalent cations directly affect the conductance of excised patches of rod photoreceptor membrane. *Science* **236**, 1674–1678 (1987).
12. Colamartino, G., Menini, A. & Torre, V. Blockage and permeation of divalent cations through the cyclic GMP-activated channel from tiger salamander retinal rods. *J. Physiol. (Lond.)* **440**, 189–206 (1991).
13. Zimmerman, A. L. & Baylor, D. A. Cation interactions within the cyclic GMP-activated channel of retinal rods from the tiger salamander. *J. Physiol. (Lond.)* **449**, 759–783 (1992).
14. Frings, S., Seifert, R., Godde, M. & Kaupp, U. B. Profoundly different calcium permeation and blockage determine the specific function of distinct cyclic nucleotide-gated channels. *Neuron* **15**, 169–179 (1995).
15. Zufall, F., Firestein, S. & Shepherd, G. M. Cyclic nucleotide-gated ion channels and sensory transduction in olfactory receptor neurons. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 577–607 (1994).
16. Armstrong, C. M. & Taylor, S. R. Interaction of barium ions with potassium channels in squid giant axons. *Biophys. J.* **30**, 473–488 (1980).
17. Armstrong, C. M., Swenson, R. P. Jr & Taylor, S. R. Block of squid axon K channels by internally and externally applied barium ions. *J. Gen. Physiol.* **80**, 663–682 (1982).
18. Jiang, Y. & MacKinnon, R. The barium site in a potassium channel by X-ray crystallography. *J. Gen. Physiol.* **115**, 269–272 (2000).
19. Heginbotham, L., Kolmakova-Partensky, L. & Miller, C. Functional reconstitution of a prokaryotic K⁺ channel. *J. Gen. Physiol.* **111**, 741–749 (1998).
20. Noskov, S. Y., Berneche, S. & Roux, B. Control of ion selectivity in potassium channels by electrostatic and dynamic properties of carbonyl ligands. *Nature* **431**, 830–834 (2004).
21. Haynes, L. & Yau, K. W. Cyclic GMP-sensitive conductance in outer segment membrane of catfish cones. *Nature* **317**, 61–64 (1985).
22. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
23. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 1772–1779 (2002).
24. Collaborative Computational Project, Number 4, The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **50**, 760–763 (1994).
25. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, S. O. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
26. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921 (1998).
27. Heginbotham, L., LeMasurier, M., Kolmakova-Partensky, L. & Miller, C. Single *Streptomyces lividans* K⁺ channels: functional asymmetries and sidedness of proton activation. *J. Gen. Physiol.* **114**, 551–560 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. MacKinnon for discussion and critical review of the manuscript. Use of the Argonne National Laboratory Structural Biology Center beamlines at the Advanced Photon Source was supported by the US Department of Energy, Office of Energy Research. We thank the beamline staff for assistance in data collection. This work was supported by grants from the David and Lucile Packard Foundation (to Y.J.) and the Searle Scholars Program (to Y.J.).

Author Contributions S.Y. and A.A. contributed equally to this work. S.Y. helped with the structure determination and A.A. performed the ⁸⁶Rb flux assay.

Author Information Atomic coordinates of the Na⁺ and K⁺ complexes of the NaK channel have been deposited in the Protein Data Bank with accession numbers of 2AHY and 2AHZ, respectively. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.J. (youxing.jiang@utsouthwestern.edu).

The candidate

Washington 2.0

Jack McDevitt

The high and low points of my career came on the same night: when we beat George Washington, and Peter Pollock returned to the White House for a second term.

Well, okay. It wasn't really Washington; it was an artificial intelligence programmed to behave like Washington. But a lot of people got confused. When you've been in politics as long as I have, you know how easily people get confused.

Fortunately,

President Pollock's numbers were down, but the Democratic candidate put everybody to sleep. So we knew it would be close. Then Washington showed up. He was a software package developed at the University of Georgia to play the part of the first president in seminars. He was so believable that somebody at the school put him on a local radio show, and next thing he was a national phenomenon: people were desperate for a candidate they could believe in. The general gave an interview to the *Florida Times-Union*, the wire services picked it up, and, by God, he did sound like George Washington.

I was running the Pollock campaign, and we all had a good laugh when they tried to put Washington on the ballot in Georgia. The Democrats tried to block it. Candidates have to be born, they said, and have to be at least 35 years old.

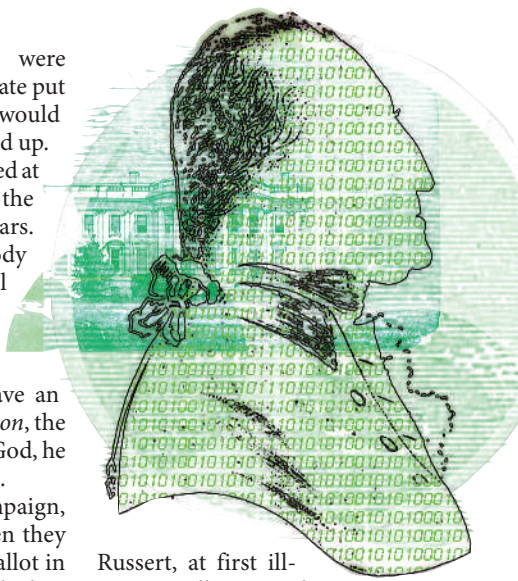
We could have stopped it then. But if Washington got into the general election, he'd pull votes from the Democrats: we knew our base wasn't going to support a candidate who wasn't even human. So I called in some favours and the Supreme Court ruled they could find no reason to suppose he was not a 'Washington-equivalent'. He was therefore clearly well past the minimum age limit. As to the requirement he be born, the software had been written in Georgia, and the meaning of 'born', said the court, is not limited to biological events.

I watched Washington on cable, and he was persuasive. He didn't like frivolous spending; didn't like unaffordable medications; didn't like corruption. I thought he came across as wooden, and maybe a trifle stern. Americans, I thought, don't like being lectured.

They could simply have done the whole thing electronically, but somebody in his campaign was smarter: he was housed in a

Coreolis 5000, and they dutifully set it on a table along with a screen which provided an animated image from the Gilbert Stuart portrait, except they'd cut the General's hair and put him in a business suit.

By midsummer he was making the rounds of the network talk shows. The week before he made his first appearance on *Meet the Press*, he passed the Democratic candidate and moved into the runner-up spot. The liberal media decided the Democratic candidate was a lost cause.



Russert, at first ill-at-ease talking to the machine, warmed to him. "Are you really George Washington?" he asked.

"The man's dead," said Washington. "But I'm everything he was."

Russert asked about the Intervention, which had become another of those endless wars. "We intended the nation to lead by example," the General said. "We would not willingly have plunged into the affairs of others. Keep your own house in order. Do it competently, and the world will follow."

We realized, belatedly, that we were in a race. After his appearance with Jon Stewart, there was no longer any doubt. "I would prefer," he told the vast audience, "that you not vote for me. And I'll tell you why: people should be governed by other people, not by software. If the voters insist, I will do my best. But I fear the long-term potential."

So we went after him: doesn't want the job. And we looked at his record: do we really want a former slave-owner in the White House?

We knew we couldn't touch him on

national security, but we demanded to know where he stood on the issues. "What about Roe vs Wade?"

"Put it aside for now," he said. "At the moment, we have bigger problems." We got some of our base back on that one.

"Gay marriage?"

"I cannot see that anyone is harmed. We should be careful about codifying moral strictures. They change too easily."

We got some more of our people back. We talked about Frankenstein. This appealed to voters so we kept hitting it. Vote for People, we said. We found a few physicists who were willing to say publicly that an artificial intelligence could develop a glitch: would you trust the Black Box in the hands of a computer?

We held on. We were still holding at 2:00 a.m. election night, when we went down to the last district in Indiana, but we took the state by a few hundred votes and that put us over the top.

Pollock went on TV after Washington conceded. He said how we'd saved the nation from a hardware conspiracy. (He tends to say things like that when he gets off script.) When it was over, he took me aside to express his appreciation. A Rainbow 360 rested on the coffee table. "We saved the country, Will," he said. "We'll get legislation to bar the things from holding office. Otherwise, I guess, they'll trot out Abe Lincoln next time."

"Yes," I said. "And congratulations, Mr President." It meant four more years for me too, as chief political adviser.

"No. It's not on the cards, Will." He looked almost genuinely pained. "We have to look to the future."

That was a shock. "What do you mean, sir?"

"It was a near thing, this election. We miscalculated our opponent's strength. I mean, incumbent president and all. It should have been easy."

"But—?"

"I need someone who won't be taken by surprise."

I was trying not to let my anger show. "Who did you have in mind, sir?"

He smiled at the Rainbow 360. "Will — meet Karl Rove."

Jack McDevitt is the author of 12 science-fiction novels and more than 60 stories. His *Omega* won the John W. Campbell Memorial Award for best novel in 2004. He has won numerous other awards and has been a consistent Nebula finalist. His current novel is *Seeker*. He lives in Georgia, USA.